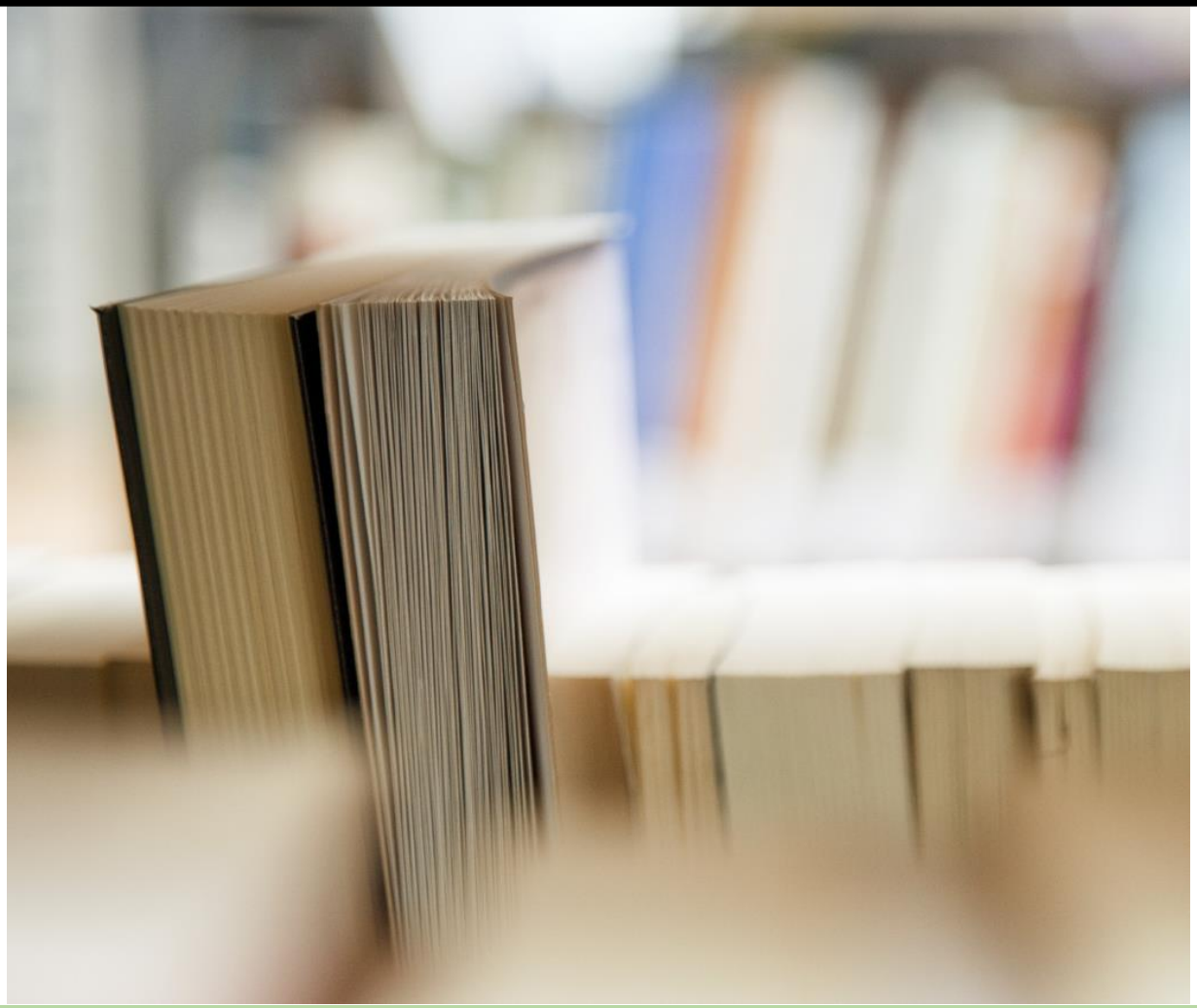
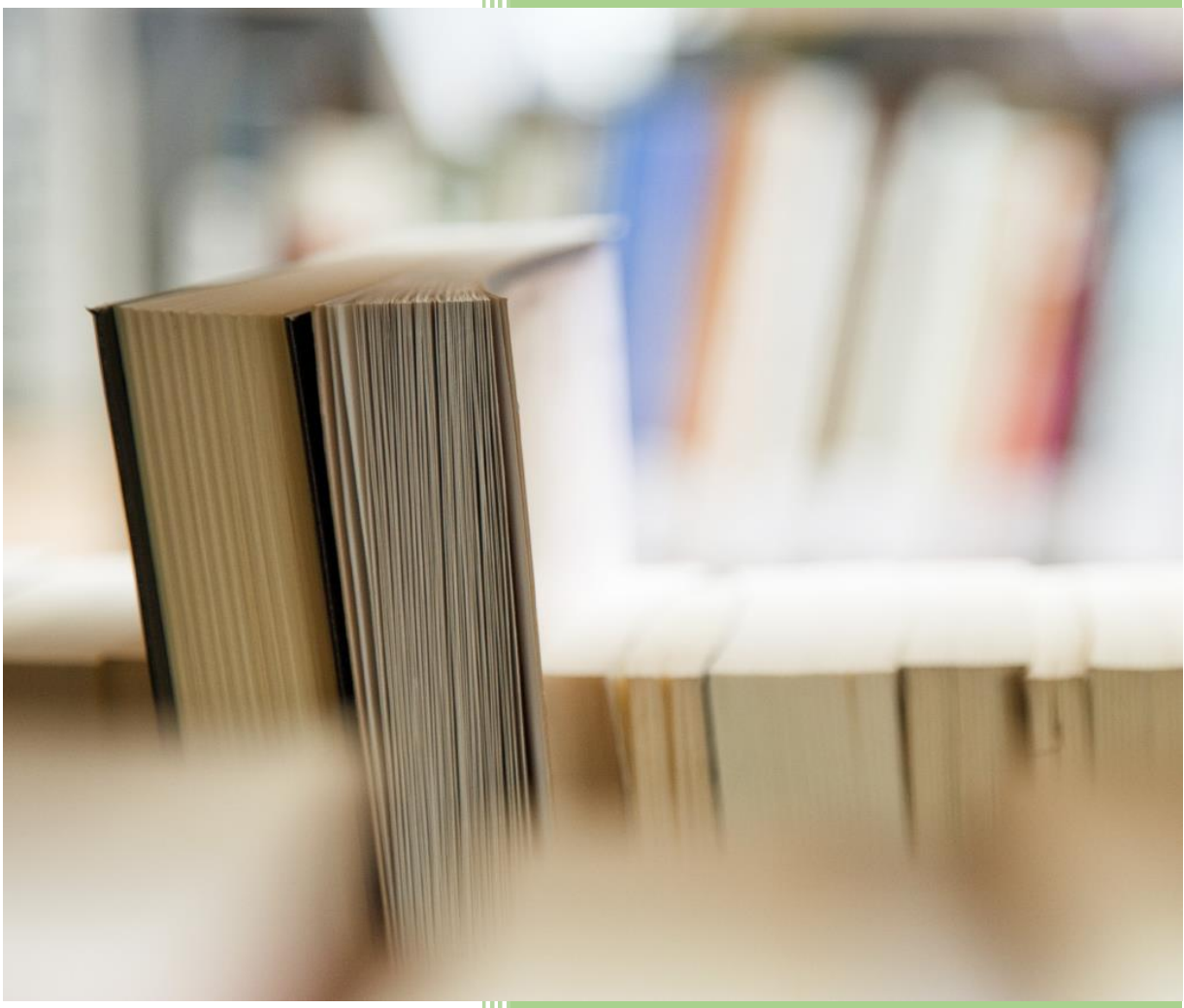
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ
ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΑΡΑΙΑ ΟΙΚΟΝΟΜΕΤΡΙΚΑ ΜΟΝΤΕΛΑ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ



ΑΝΔΡΕΑΣ ΠΥΡΙΛΛΗ
ΕΠΙΒΛΕΠΩΝ: Κος. ΧΡΗΣΤΟΣ ΚΟΥΚΟΥΒΙΝΟΣ
2016, ΑΘΗΝΑ

HIGH DIMENSIONAL SPARSE ECONOMETRIC MODELS



ANDREAS PYRILLI

SEMFE, NTUA

2016, ATHENS

Pr. CHRISTOS KOUKOUVINOS

ΠΕΡΙΛΗΨΗ

Με την πάροδο του χρόνου η στατιστική ανάλυση και η επιλογή των “σημαντικών” μεταβλητών σε δεδομένα υψηλών διαστάσεων γίνεται ολοένα και πιο συχνή σε διάφορους τομείς των θετικών και των ανθρωπιστικών επιστημών. Παρουσιάζεται ιδιαίτερο ενδιαφέρον όχι μόνο επειδή περιλαμβάνει ελκυστικές εφαρμογές, αλλά και επειδή ένα μεγάλο μέρος της στατιστικής ανάλυσης θα πρέπει να επανεξεταστεί. Η διπλωματική αυτή εστιάζει την προσοχή της στην κατηγορία των ποινικοποιημένων μεθόδων που χρησιμεύουν για την επιλογή των μεταβλητών και στα αραιά οικονομετρικά μοντέλα υψηλών διαστάσεων. Συνάμα, γίνεται μια ανάλυση στα διάφορα θεωρήματα και αποτελέσματα που περιστρέφονται γύρω από την συγκεκριμένη θεματική ενότητα. Στην ουσία η εργασία αρχικά επικεντρώνεται στις μεθόδους LASSO ενώ στην πορεία ασχολείται με τον εκτιμητή Dantzig Selector.

Συγκεκριμένα, στο 1ο κεφάλαιο γίνεται μια εισαγωγή στο θέμα των υψηλών διαστάσεων και τις προκλήσεις που το περιτριγυρίζουν έτσι ώστε να βοηθήσει τον αναγνώστη να κατανοήσει την έννοια της υψηλής διαστατικότητας · κυρίως στον τομέα της οικονομίας. Ακολουθεί ανάλυση του αραιού μοντέλου υψηλών διαστάσεων, η συνθήκη ASM καθώς και κάποια παραδείγματα για την εφαρμογή των θεωρητικών αποτελεσμάτων. Στη συνέχεια, το 2ο κεφάλαιο μας εισάγει στον κόσμο των ποινικοποιημένων μεθόδων και ειδικότερα στις μεθόδους LASSO που αποτελούν και ένα από τα βασικότερα κλειδιά της εργασίας. Παρουσιάζονται οι διάφορες παραλλαγές της LASSO (Iterated LASSO, Square-root LASSO, Post LASSO, Adaptive LASSO) και ότι περιστρέφεται γύρω από αυτές. Το 3ο κεφάλαιο αποτελεί απόρροια του προηγούμενου μιας και γίνεται παρουσίαση των αποτελεσμάτων εκτίμησης για τα αραιά μοντέλα υψηλών διαστάσεων σχετικά με τις μεθόδους LASSO και Post LASSO. Ακολούθως, το 4ο κεφάλαιο ασχολείται με τις βοηθητικές μεταβλητές και πως αυτές συνδέονται με την υψηλή διαστατικότητα. Στο 5ο κεφάλαιο αναλύεται ο εκτιμητής Dantzig Selector και ότι σχετίζεται με αυτόν · μεταξύ των οποίων είναι η ομοιόμορφη αρχή της Αβεβαιότητας, Θεωρήματα σχετικά με τα αποτελέσματα εκτίμησης τους και ο Gauss-Dantzig Selector. Έπειτα, το 6ο κεφάλαιο συνεχίζει να ασχολείται με τον εκτιμητή Dantzig Selector με την διαφορά ότι επικεντρώνεται στην δράση που έχει κυρίως στα μερικώς γραμμικά μοντέλα. Τέλος, στο 7ο κεφάλαιο εκτελείται μια εφαρμογή στο στατιστικό πρόγραμμα R με τη βοήθεια του στατιστικού πακέτου “flare”.

Λέξεις κλειδιά: **Υψηλή διαστατικότητα, Αραιό οικονομετρικό μοντέλο υψηλών διαστάσεων, Ποινικοποιημένη μέθοδος , LASSO, Dantzig Selector.**

ABSTRACT

As time goes by, statistical analysis and the selection of significant variables in high dimensional data become more frequent in several fields of Applied and Human Sciences. This interest emanates, not only from the attractive applications, but also from the need for a revision of a great part of statistical analysis. This thesis focuses on the category of penalization methods which can be used in the selection of the variables and the high dimensional sparse econometric models. Also, analysis is done on several theorems and results, which rotate around the specific thematic unit. Fundamentally, the paper, at first, focuses on the LASSO methods while, subsequently deals with the εκτιμητή Dantzig Selector.

Specifically, the 1st chapter introduces the subject of high dimensional and the challenges that surround it in order for the reader to understand the concept of high dimensionality; especially in the field of economics. Following, analysis of high dimensional sparse model, condition ASM but, also, some examples of applications of the theoretical results. Thereafter, the 2nd chapter introduces us to the world of penalization methods and mainly to the LASSO methods, which consist a key aspect of the paper. Different variations of LASSO (Iterated LASSO, Square-root LASSO, Post LASSO, Adaptive LASSO) are presented and their surroundings. The 3rd chapter is a corollary of the previous one as a presentation of the estimation results is done for the high dimensional sparse models in respect of the LASSO methods and Post LASSO. Thereupon, the 4th chapter deals with the instrumental variables and how they are correlated with the high dimensionality. In the 5th chapter the estimator Dantzig Selector is being analyzed and everything that relates to it; among which is the Uniform Uncertainly Principle, theorems concerning their estimation results and the Gauss-Dantzig Selector. Afterwards, the 6th chapter continues to deal with the estimator Dantzig Selector, with the difference that it focuses upon its effect, primarily, on the partially linear models. In conclusion, in the 7th chapter an application is executed on the R statistical suite with the help of the package “flare”.

Key words: **High dimensionality, High dimensional sparse econometric model, Penalization method, LASSO, Dantzig Selector.**

ΕΥΧΑΡΙΣΤΙΕΣ

Η εκπόνηση της συγκεκριμένης διπλωματικής εργασίας έγινε υπό την επίβλεψη του Καθηγητή του Εθνικού Μετσόβιου Πολυτεχνείου κύριου Χρήστου Κουκουβίνου · τον οποίο και θα ήθελα να ευχαριστήσω θερμά για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο καθώς και για την σημαντική του καθοδήγηση καθ' όλη την διάρκεια της εκπόνησης.

Επίσης, οφείλω ένα μεγάλο ευχαριστώ στην υποψήφια διδάκτωρ Αγγελική Λάππα για την πολύτιμη βοήθεια της και την άψογη συνεργασία που είχαμε.

Ιδιαίτερες ευχαριστίες θέλω να εκφράσω προς την οικογένεια μου για την διαρκή στήριξη που μου πρόσφεραν κατά την διάρκεια των σπουδών μου και όχι μόνο.

Πυρίλλη Ανδρέας

Εθνικό Μετσόβιο Πολυτεχνείο,

Σχολή Εφαρμοσμένων Μαθηματικών

και Φυσικών Επιστημών

Αθήνα, 2016

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΕΧΟΜΕΝΑ ΠΙΝΑΚΩΝ

ΠΙΝΑΚΑΣ ΣΥΝΤΟΜΟΓΡΑΦΙΩΝ

VAR	Vector Autoregressive models	Μοντέλο Αυτοπαλινδρομι-κού Διανύσματος
FAVAR	Factor-Augmented Vector Autoregressive	Εκτιμώμενοι Παράγοντες Αυτοπαλινδρομικού Διανύσματος
LASSO	Least Absolute Shrinkage and Selection Operator	-
ASM	Approximate Sparse Model	Προσεγγιστικό Αραιό Μοντέλο
OLS	Ordinary Least Squares	Εκτιμητής Ελάχιστων Τετραγώνων
PLS	Penalized Least Squares	Ποινικοποιημένα ελάχιστα τετράγωνα
AIC	Akaike Information Criterion	Κριτήριο Πληροφορίας Akaike
BIC	Bayesian Information Criterion	Κριτήριο Πληροφορίας Bayesian
GDP	Gross Domestic Product	Ακαθάριστο Εγχώριο Προϊόν
IV	Instrumental Variables	Βοηθητικές Μεταβλητές
TSLS	Two-Stage Least Squares	Εκτιμητής Ελάχιστων Τετραγώνων σε Δύο Στάδια
UUP	Uniform Uncertainly Principle	Ομοιόμορφη Αρχή της Αβεβαιότητας
DS	Dantzig Selector	Επιλογέας Dantzig
MSE	Mean Square Error	Μέσο Τετραγωνικό Σφάλμα
LARS	Least-Angle Regression	Παλινδρόμηση ελάχιστης γωνιάς
DASSO	Dantzig Selector with Sequential Optimization	Επιλογέας Dantzig και Ακολουθιακή Βελτιστοποίηση

ΚΕΦΑΛΑΙΟ 1

ΑΡΑΙΑ ΜΟΝΤΕΛΑ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ ΣΤΗΝ ΟΙΚΟΝΟΜΕΤΡΙΑ

1.1 Εισαγωγή

Ένα μεγάλο μέρος της στατιστικής ασχολείται με την κατασκευή και ανάλυση στατιστικών μοντέλων (προτύπων). Με τον όρο μοντέλο εννοούμε τη μορφή της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών. Η ανάλυση παλινδρόμησης (*regression analysis*) είναι μια ευρέως χρησιμοποιημένη στατιστική τεχνική μοντελοποίησης για την έρευνα της συσχέτισης μεταξύ μιας εξαρτώμενης (*dependent / response*) μεταβλητής και μίας ή περισσότερων ανεξάρτητων (*independent / regressor / predictor*) μεταβλητών σε ένα συγκεκριμένο σύνολο δεδομένων. Σε πειραματικές έρευνες, η ανεξάρτητη μεταβλητή x είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή να καθορίσουμε τις τιμές της. Από την άλλη, η εξαρτημένη μεταβλητή y είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές.

Στην πιο απλή περίπτωση του γραμμικού μοντέλου ο σκοπός μας είναι η προσαρμογή μιας ευθείας γραμμής, η οποία θα επεξηγεί όσο το δυνατό καλύτερα τη συμπεριφορά των δεδομένων μας. Μια τέτοια ευθεία θα έχει τη μορφή:

$$E(y|x) = E(y_x) = \beta_0 + \beta_1 x \qquad (1.1)$$

όπου τα β_0 και β_1 αποτελούν τις παραμέτρους του μοντέλου ή αλλιώς τους συντελεστές παλινδρόμησης (*coefficients*). Με λίγα λόγια, η παραπάνω σχέση περιγράφει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής y_x , όταν η ανεξάρτητη μεταβλητή πάρει την τιμή x. Η μεταβλητή y θεωρείται ότι είναι μια τυχαία μεταβλητή, ενώ αντιθέτως η μεταβλητή x θεωρείται μη στοχαστική.

Βεβαίως υπάρχουν πολύ πιο σύνθετες περιπτώσεις από την παραπάνω όπου είτε πιθανώς το μοντέλο να είναι μη γραμμικό ή να υπάρχουν περισσότερες από μια εξαρτημένες μεταβλητές x, που επηρεάζουν την συμπεριφορά του μοντέλου, και κατά συνέπεια περισσότεροι παράμετροι β. Εντούτοις ένα χαρακτηριστικό γνώρισμα που παρουσιάζεται σε πολλά σύγχρονα προβλήματα είναι ο αριθμός των παραμέτρων, p, να είναι πολύ μεγαλύτερος από το μέγεθος του δείγματος, n, πράγμα που δεν παρουσιαζόταν στις συνηθισμένες περιπτώσεις. Όταν η διάσταση p είναι υπερβολικά υψηλή, συχνά θεωρείται ότι μόνο ένας μικρός αριθμός μεταβλητών μεταξύ των επεξηγηματικών μεταβλητών $x_1, \dots, x_p$ συμβάλλει στην απόκριση,γεγονός που οδηγεί στη σποραδικότητα του διανύσματος παραμέτρων β. Κατά συνέπεια, η επιλογή μεταβλητών διαδραματίζει εξέχοντα ρόλο στη στατιστική μοντελοποίηση δεδομένων μεγάλων διαστάσεων (*high dimensional data*).

Η ανάλυση δεδομένων υψηλής διάστασης γίνεται όλο και πιο συχνή και σημαντική σε διάφορους τομείς των θετικών επιστημών, της μηχανικής και των ανθρωπιστικών επιστημών, που κυμαίνονται από τη γονιδιωματική και τις επιστήμες υγείας στην οικονομία και τη μηχανική μάθηση. Είναι ιδιαίτερα ενδιαφέρουσα όχι μόνο επειδή περιλαμβάνει ενδιαφέρουσες εφαρμογές, αλλά επειδή ένα μεγάλο μέρος της παραδοσιακής στατιστικής ανάλυσης θα πρέπει να επανεξεταστεί.

1.1.2 Προκλήσεις της υψηλής διαστατικότητας

Η υψηλή διαστατικότητα θέτει πολυάριθμες προκλήσεις στη στατιστική θεωρεία, στις μεθόδους και στις εφαρμογές των συγκεκριμένων προβλημάτων. Παραδείγματος χάριν, στο γραμμικό μοντέλο της παλινδρόμησης με μεταβλητή θορύβου (*noise variable*) σ^2, όταν η διάσταση του μοντέλου p είναι συγκρίσιμη ή και μεγαλύτερη από το μέγεθος του δείγματος n, τότε ο εκτιμητής ελάχιστων τετραγώνων (*ordinary least squares estimator*) δεν συμπεριφέρεται καλά ή και ακόμη δεν είναι πλέον μοναδικός · εξαιτίας της ιδιομορφίας που παρουσιάζει ο πίνακας σχεδιασμού (*design matrix*) X.

Το μοντέλο παλινδρόμησης που φτιάχνεται από όλες τις ανεξάρτητες μεταβλητές έχει συνήθως σφάλμα πρόβλεψης της τάξεως του $(1 + p/n)^{1/2} \cdot \sigma$ όταν το $p \leq n$, έναντι του $(1 + s/n)^{1/2} \cdot \sigma$ όταν υπάρχουν μόνο s ανεξάρτητες μεταβλητές. Αυτή η εξέλιξη αντανακλά δύο γνωστά φαινόμενα στην μοντελοποίηση υψηλών διαστάσεων. Την συγγραμικότητα (*collinearity*) ή τους πλασματικούς συσχετισμούς (*spurious correlations*) και τη συσσώρευση θορύβου.

1.1.3 Τα δεδομένα υψηλών διαστάσεων στο κλάδο της οικονομίας

Όπως προαναφέρθηκε, τα τελευταία χρόνια τα μοντέλα υψηλών διαστάσεων έχουν αποκτήσει ιδιαίτερη σημασία σε διάφορους επιστημονικούς κλάδους · και προ πάντων σε αυτό της οικονομίας. Ένα παράδειγμα της υψηλής διαστατικότητας έχει να κάνει με το μοντέλο του αυτοπαλινδρομικού[1] διανύσματος (*Vector Autoregressive models, VAR*) των Sims (1980), Stock και Watson (2001) το οποίο αποτελεί το κλειδί για να αναλύσει κανείς την από κοινού εξέλιξη των μακροοικονομικών χρονοσειρών και να προσφέρει έτσι πολλές δομικές πληροφορίες. Εξαιτίας του γεγονότος ότι ο αριθμός των παραμέτρων αυξάνεται δραματικά σε σχέση με το μέγεθος του μοντέλου, συνήθως το τυποποιημένο *VAR* δεν περιλαμβάνει περισσότερες από δέκα μεταβλητές. Ωστόσο, οι επιστήμονες είναι σε θέση να παρατηρούν εκατοντάδες σειρές δεδομένων. Για να εμπλουτίσουν το σύνολο των πληροφοριών του μοντέλου μια ομάδα ερευνητών (Bernarke et al. 2005) πρότειναν την αύξηση των τυποποιημένων *VAR* με τους εκτιμώμενους παράγοντες (*Factor-Augmented Vector Autoregressive, FAVAR*) για την μέτρηση των αποτελεσμάτων της νομισματικής πολιτικής.

Ένα άλλο παράδειγμα σχετικά με τα δεδομένα υψηλών διαστάσεων είναι τα μεγάλα δεδομένα του πίνακα τιμών ενός σπιτιού. Ας μελετήσουμε την περίπτωση μιας μεγάλης χώρας όπως είναι οι Η.Π.Α. Για να ενσωματωθούν τα αποτελέσματα του αντιπροσωπευτικού δείγματος θα πρέπει να έχουμε κατά νου

[1] Αυτοπαλινδρόμηση (*autoregression*): Έχει να κάνει με τα προβλήματα χρονοσειρών στα οποία μας ενδιαφέρει κυρίως η εξάρτηση που μπορεί να έχει η τυχαία μεταβλητή X_t από τις προηγούμενες τυχαίες μεταβλητές $X_{t-1}, X_{t-2}, \ldots$ για κάθε χρονική στιγμή t.

ότι η τιμή σε μια κομητεία μπορεί να εξαρτάται από πολλές άλλες κομητείες και ειδικότερα από τις κομητείες που είναι γειτονικές με την συγκεκριμένη. Δεδομένου, ότι τέτοιος συσχετισμός είναι άγνωστος, αρχικά η εξίσωση παλινδρόμησης μπορεί να περιλαμβάνει περίπου 1000 κομητείες των Η.Π.Α, γεγονός που όπως γίνεται αντιληπτό κάνει αδύνατη την εκτίμηση των ελάχιστων τετραγώνων του μοντέλου. Μια προτεινόμενη τεχνική για την μείωση των διαστάσεων του μοντέλου είναι η επιλογή μεταβλητών (*variable selection*) έτσι ώστε το μοντέλο να περιέχει μόνο τις μεταβλητές που είναι σημαντικές. Οι επιστήμονες της στατιστικής και της οικονομετρίας δημιούργησαν κάποιους αλγορίθμους για να μπορούν ταυτόχρονα να επιλέγουν τις σχετικές μεταβλητές (*relevant variables*) και να εκτιμούν αποτελεσματικά τις παραμέτρους. Οι τεχνικές για την επιλογή μεταβλητών, έχουν χρησιμοποιηθεί ευρέως τόσο στην κατασκευή χρηματοοικονομικών χαρτοφυλακίων (*financial portfolio*) όσο και στα μοντέλα πιστωτικού κινδύνου (*credit risk models*).

Πρόσφατα έχουν προταθεί κάποιες νέες μέθοδοι σχετικά με την επιλογή των μεταβλητών , όπως είναι η LASSO (Tibshirani (1996)), η παλινδρόμηση bridge (Frank and Friedman (1993) , Fu (1998), Knight και Fu (2000)), το ελαστικό δίκτυο (*elastic net*, Zou και Hastie (2005)) , ο αραιός εκτιμητής ενός βήματος (the one-step sparse estimator, Zou και Li (2008)), η παλινδρόμηση ελάχιστης γωνιάς (LARS algorithm, Efron et al. (2004)) και ο επιλογέας Dantzig (Dantzig selector, Candes και Tao (2007), Bickel et al. (2009), Meinshausen et al. (2007), James and Radchenko (2009), Dicker και Lin (2009)). Στη παρούσα διπλωματική θα ασχοληθούμε με τη μέθοδο LASSO και τον Dantzig selector.

1.2 Αραιό μοντέλο υψηλών διαστάσεων στην οικονομετρία

Σε αυτή την ενότητα γίνεται μια σύντομη παρουσίαση στα γραμμικά αραιά μοντέλα υψηλών διαστάσεων (*High Dimensional Sparse Models*) εφαρμοσμένα στο κλάδο της οικονομετρίας. Το μοντέλο αυτό, εξ' ορισμού κάνει χρήση της υπόθεσης της σποραδικότητας (*sparsity*), χρησιμοποιώντας μικρότερο αριθμό εκ' των διαθέσιμων ανεξάρτητων μεταβλητών. Συγκεκριμένα, υποθέτουμε ότι μόνο $s \ll n$ από αυτές τις μεταβλητές θα είναι σημαντικές για την κατασκευή

του ζητούμενου μοντέλου. Αυτό όπως θα δούμε σε επόμενα κεφάλαια γίνεται με την βοήθεια ποινικοποιημένων μεθόδων (*penalization methods*).

Καταρχήν, έστω ότι έχουμε το ακόλουθο παραμετρικό μοντέλο παλινδρόμησης :

$$y_i = x_i{}'\beta_o + \varepsilon_i \ , \ \text{με } \beta_o \in R^p, i = 1,2,\dots,n \qquad (1.2)$$

όπου y_i ως γνωστόν είναι οι παρατηρήσεις της μεταβλητής απόκρισης, ενώ τα x_i αντιστοιχούν στις παρατηρήσεις των επεξηγηματικών (ανεξάρτητων) μεταβλητών p-διαστάσεων. Για την εξέταση του πιο πάνω μοντέλου κάνουμε χρήση της υπόθεσης ότι η κατανομή των τυχαίων σφαλμάτων, ε_i, είναι η Κανονική (γνωστή και ως Γκαουσιανή κατανομή) · δηλαδή $\varepsilon_i \sim N(0, \sigma^2)$.

1.2.1 Σποραδικότητα

Η αραιή μοντελοποίηση έχει χρησιμοποιηθεί ευρέως για να εξετάσει την υψηλή διαστατικότητα. Η βασική ιδέα είναι ότι το παραμετρικό διάνυσμα p-διαστάσεων β_o είναι αραιό με πολλές συνιστώσες να είναι ακριβώς μηδέν ή και αμελητέες, ενώ κάθε μη μηδενική συνιστώσα του θα αντιπροσωπεύει τη συμβολή μιας σημαντικής ανεξάρτητης μεταβλητής (*predictor*). Αναλυτικότερα, θα έχει μόνο $s < n$ μη μηδενικές συνιστώσες με φορέα (*support*) του μοντέλου να ορίζεται ο $T = support(\beta_0) \subset \{1,2,\dots,p\}$ · το λεγόμενο και πραγματικό μοντέλο (*true model*).

Τέτοια υπόθεση είναι κρίσιμη για τον προσδιορισμό του πραγματικού αραιού μοντέλου · ειδικά όταν δίνεται ένα σχετικά μικρό μέγεθος του δείγματος. Αν και η έννοια της σποραδικότητας δίνει αφορμή για μεροληπτική εκτίμηση, σε γενικές γραμμές έχει αποδειχθεί αρκετά αποτελεσματική σε πολλές εφαρμογές. Επίσης, η σποραδικότητα θα πρέπει να γίνει κατανοητή σε μια ευρύτερη έννοια, ως μειωμένη πολυπλοκότητα. Παραδείγματος χάριν, η υπόθεση της σποραδικότητας γίνεται από οικονομικούς αναλυτές σε περιπτώσεις που θεωρούν ότι

το αποτέλεσμα θα μπορούσε να γίνει καλύτερο αν χρησιμοποιούσαν ένα μικρότερο αριθμό παραγόντων σε σχέση με το δείγμα όπου η ταυτότητα τους πιθανόν να είναι άγνωστη.

Επομένως, αναζητούμε μια λύση του β_o με $s \ll n$ στοιχεία να είναι μη μηδενικά, των οποίων η ταυτότητα παραμένει άγνωστη. Συγχρόνως, υποθέτουμε ότι $p = p_n$ αυξάνεται προς το άπειρο καθώς αυξάνεται το n, και $s = s_n$ αυξάνεται καθώς αυξάνεται το n , αν και θα απαιτηθεί και η χρήση του $slogp = o(n)$.

Στην ουσία το αραιό μοντέλο είναι μια γενίκευση του κλασσικού παραμετρικού γραμμικού μοντέλου, αφήνοντας τις ταυτότητες T των σχετικών μεταβλητών να είναι άγνωστες. Η συγκεκριμένη γενίκευση είναι αρκετά βολική μιας και σε πολλές περιπτώσεις είναι δύσκολο να γνωρίζουμε τις ταυτότητες των σχετικών ανεξάρτητων μεταβλητών.

Το μοντέλο που παρουσιάστηκε πιο πάνω είναι το απλό και στη θεωρία μας επιτρέπει να εφαρμόσουμε προσεγγιστικά την υπόθεση της σποραδικότητας. Παρόλα αυτά δεν συνάδει με την πραγματικότητα, υπό την έννοια ότι προϋποθέτει ακριβής αραιότητα ή ότι μετά το πέρας του υπολογισμού των s βασικών ανεξάρτητων μεταβλητών, το σφάλμα στη προσέγγιση της συνάρτησης παλινδρόμησης θα είναι μηδέν.

Για τον λόγο αυτό, θα παραβλέψουμε το πιο πάνω μοντέλο και θα ασχοληθούμε με ένα πιο γενικό, αραιό προσεγγιστικό ή μη παραμετρικό μοντέλο. Στο συγκεκριμένο μοντέλο, όλες οι ανεξάρτητες μεταβλητές πιθανώς να έχουν μη μηδενική συνεισφορά στην συνάρτηση παλινδρόμησης, αλλά δεν θα είναι τίποτα παραπάνω από τις s άγνωστες ανεξάρτητες μεταβλητές που απαιτούνται για την προσεγγιστική μορφή της συνάρτησης παλινδρόμησης με επαρκή βαθμό ακρίβειας.

1.2.2 Συνθήκη Προσεγγιστικού Αραιού Μοντέλου

Στην παρούσα υποενότητα θα αναφερθούμε στην Συνθήκη Προσεγγιστικού Αραιού Μοντέλου (*Approximate Sparse Model, ASM*).

Έστω το σύνολο των δεδομένων $\{(y_i, z_i), i = (1,2, \dots, n)\}$ όπου για κάθε n υπακούει στο μοντέλο παλινδρόμησης

$$y_i = f(z_i) + \varepsilon_i \ , \quad \varepsilon_i \sim N(0, \sigma^2), \ i = 1,2, \dots, n \tag{1.3}$$

όπου y_i είναι η μεταβλητή εξόδου , z_i είναι το k_z- διάνυσμα των στοιχειωδών ανεξάρτητων μεταβλητών, $f(z_i)$ είναι η πραγματική συνάρτηση παλινδρόμησης που πιθανόν να είναι μη γραμμική, ε_i είναι τα τυχαία σφάλματα με κανονική κατανομή.

Ορίζουμε $x_i \coloneqq P(z_i)$, όπου $P(z_i)$ είναι ένα διάνυσμα με διάσταση $p = p_n$, το οποίο διάνυσμα αποτελείται από πιθανούς τεχνικούς μετασχηματισμούς των z_i, συμπεριλαμβανομένης και της σταθεράς, και $f_i = f(z_i)$. Ειδικότερα, το διάνυσμα $x_i = P(z_i)$ θα περιλαμβάνει είτε πολυωνυμικούς είτε spline μετασχηματισμούς των αρχικών ανεξάρτητων μεταβλητών z_i.

Οι τιμές $x_1, \dots, x_n$ αντιμετωπίζονται ως σταθερές και είναι κανονικοποιημένες, γεγονός που σημαίνει ότι ισχύει $\hat{\sigma}_j^2 = E_n[{x_{ij}}^2] = 1$ για κάθε $j = 1, \dots, p$.

Η προσέγγιση της αραιότητας εισάγεται με την βοήθεια της συνάρτησης παλινδρόμησης $f(z_i)$. Ειδικότερα, αναφέρεται ότι υπάρχουν β_o για τα οποία ισχύουν τα παρακάτω:

$$f(z_i) = {x_i}'\beta_o + r_i \ , \qquad \|\beta_o\|_o \leq s \ , \qquad c_s \coloneqq \{E_n[{r_i}^2]\}^{\frac{1}{2}} \leq K\sigma\sqrt{\frac{s}{n}} \tag{1.4}$$

όπου r_i είναι τα προσεγγιστικά σφάλματα , $s = s_n = o\left(\frac{n}{logp}\right)$ και η σταθερά K είναι ανεξάρτητη από το μέγεθος του δείγματος n.

1.2.2.1 Αποτελέσματα της συνθήκης ASM

Σχετικά με το μοντέλο, εξετάζουμε την περίπτωση του σταθερού σχεδιασμού, η οποία καλύπτει μια τυχαία δειγματοληψία ως μια ειδική περίπτωση κατά την οποία τα $x_1, \dots, x_n$ θα αντιπροσωπεύουν την υλοποίηση αυτού του δείγματος. Η προσέγγιση της αραιότητας στο ζητούμενο μοντέλο είναι δυνατό να γίνει με παρόμοιο τρόπο με αυτό του Newey, W.K. μέσα από το άρθρο του

"*Convergence Rates and Asymptotic Normality for Series Estimators*" (1997), στο οποίο αναλύεται ότι οι πρώτοι $s = s_n$ όροι της σειράς είναι αυτοί που θα προσεγγίσουν πολύ καλά την μη παραμετρική συνάρτηση της παλινδρόμησης.

Παρόλα αυτά η περίπτωση της συνθήκης ASM που εξετάζουμε είναι γενικότερη και διαφοροποιείται στο σημείο το οποίο οι στατιστικά σημαντικοί $s = s_n$ όροι της προσέγγισης δεν είναι κατ' ανάγκη οι πρώτοι s όροι. Στην πραγματικότητα θα είναι άγνωστη η ταυτότητα τους.

Όσον αφορά την μη παραμετρική περίπτωση, θεωρούμε την ποσότητα $x_i'\beta_o$ ως ένα αραιό παραμετρικό μοντέλο το οποίο παράγει μια σχετικά καλή προσέγγιση στην πραγματική συνάρτηση της παλινδρόμησης $f(z_i) = x_i'\beta_o + r_i$ της εξίσωσης $y_i = f(z_i) + \varepsilon_i$, έτσι ώστε το r_i να είναι μικρό σχετικά με το υποτιθέμενο μέγεθος του εκτιμώμενου σφάλματος. Από την άλλη, στην παραμετρική περίπτωση, το r_i μηδενίζεται μιας και επιλέγουμε να ισχύει $x_i'\beta_o = f(z_i)$ για κάθε $i = 1, \dots, n$.

Ο πρωταρχικός μας στόχος κατά την εκτίμηση είναι η συνάρτηση παλινδρόμησης $f(z_i)$. Παρόλα αυτά στρέφουμε το ενδιαφέρον μας στην εκτίμηση της παραμετρικής συνάρτησης $x_i'\beta_o$ έτσι ώστε να προσεγγίσουμε την εκτίμηση του αρχικού προβλήματος σαν να ήταν παραμετρικό. Είναι φανερό ότι οι δυο πιο πάνω στόχοι είναι ισοδύναμοι όσων αφορά την επιλογή των προσεγγιστικών σφαλμάτων r_i να είναι μικρότερα από τα σφάλματα εκτίμησης.

Ένας τρόπος για να κατασκευαστεί αναλυτικά μια καλή προσέγγιση του μοντέλου β_o για την σχέση της συνάρτησης παλινδρόμησης (1.4) είναι να πάρουμε το β_o ως λύση του παρακάτω προβλήματος:

$$min_{\beta \in R^p} E_n[(f_i - x_i'\beta)^2] + \frac{\sigma^2 \|\beta_o\|}{n} \qquad (1.5)$$

Ονομάζουμε την σχέση (1.5) πρόβλημα *oracle* [2], και επίσης το πραγματικό μοντέλο $T = support(\beta_0)$ ορίζεται και ως μοντέλο *oracle*. Να τονισθεί ότι ισχύει ότι $s = \|\beta_o\| \leq n$.

[2] Εξ' ορισμού το πρόβλημα oracle γνωρίζει την συνάρτηση κινδύνου οποιουδήποτε εκτιμητή και έτσι μπορεί να υπολογίσει τον καλύτερο αραιό εκτιμητή ελάχιστων τετραγώνων. Υπό κάποια συνθήκη το

Με την χρήση του προβλήματος oracle (1.5) γίνεται μια εξισορρόπηση του προσεγγιστικού σφάλματος $E_n[(f(z_i) - x_i'\beta_o)^2]$ πάνω από τα σημεία σχεδιασμού με τον όρο της διασποράς $\sigma^2\|\beta_o\|/n$. Ο όρο αυτός της διασποράς, προσδιορίζεται από τον αριθμό των μη μηδενικών συντελεστών του β.

Θέτοντας $c_s{}^2 \coloneqq E_n[r_i{}^2] = E_n[(f(z_i) - x_i'\beta_o)^2]$, ορίζουμε το μέσο τετραγωνικό σφάλμα των προσεγγιστικών τιμών $f(z_i)$ από την $x_i'\beta_o$. Έτσι με αντικατάσταση, η προκύπτουσα ποσότητα $c_s{}^2 + \sigma^2 s/n$ αποτελεί την βέλτιστη τιμή του παραπάνω προβλήματος *oracle*.

Όταν έχουμε να κάνουμε με μη παραμετρικές καταστάσεις, η βέλτιστη λύση του προβλήματος *oracle* εξισορροπεί την προσέγγιση του σφάλματος με τον όρο της διασποράς δίνοντας ότι θα ισχύει $c_s{}^2 \leq K\sigma\sqrt{s/n}$. Επομένως, προκύπτει ότι $\sqrt{c_s{}^2 + \sigma^2 s/n} \leq \sigma\sqrt{s/n}$, πράγμα που σημαίνει ότι η ποσότητα $\sigma\sqrt{s/n}$ είναι ιδανική για τον ρυθμό σύγκλισης. Στην περίπτωση που μας ήταν γνωστό το μοντέλο *oracle* T, τότε θα μπορούσαμε να επιτύγχουμε αυτό το ρυθμό σύγκλισης, χρησιμοποιώντας τον εκτιμητή *oracle*, δηλαδή τον εκτιμητή ελάχιστων τετραγώνων ο οποίος είναι βασισμένος στο μοντέλο *oracle* T. Στην πραγματικότητα όμως, δεν είναι γνωστό το *oracle* μοντέλο T μιας και δεν παρατηρούμε την $f(z_i)$ και ως εκ τούτου είναι αδύνατο να λύσουμε το πρόβλημα *oracle* (1.5). Από την στιγμή που το T είναι άγνωστο όπως προαναφέρθηκε, είναι δύσκολο να επιτευχθεί ο ακριβές βαθμός σύγκλισης του *oracle*. Παρόλα' αυτά προσδοκούμε να επιτύχουμε όσο το δυνατό περισσότερο αυτό το ρυθμό.

1.2.2.2 Πρόβλημα Oracle

Υπό ορισμένες ήπιες υποθέσεις, το πρόβλημα (1.5) προκύπτει άμεσα ως το oracle πρόβλημα ελαχιστοποίησης του κινδύνου. Πράγματι, θεωρώντας ένα εκτιμητή ελάχιστων τετραγώνων (*Ordinary least squares, OLS*) $\hat{\beta}[\tilde{T}]$, ο οποίος αποκτάται χρησιμοποιώντας ένα μοντέλο $\tilde{T}$, δηλαδή με παλινδρόμηση y_i στις

πρόβλημα ελαχιστοποίησης της πρόβλεψης κινδύνου μεταξύ όλων των αραιών εκτιμητών ελάχιστων τετραγώνων είναι ισοδύναμο με το συγκεκριμένο πρόβλημα.

ανεξάρτητες μεταβλητές $x_i[\tilde{T}]$, όπου $x_i[\tilde{T}] = \{x_{ij}, j \in \tilde{T}\}$. Αυτός ο εκτιμητής παίρνει την τιμή $\hat{\beta}[\tilde{T}] = E_n\,[x_i[\tilde{T}]x_i[\tilde{T}]']^{-1}\,E_n[x_i[\tilde{T}]y_i]$.

Ο αναμενόμενος κίνδυνος αυτού του εκτιμητή $E_n E[f_i - x_i|\tilde{T}]'\hat{\beta}[\tilde{T}]]^2$ είναι ίσος με

$$min_{\beta \in R^{|\tilde{T}|}} E_n[(f_i - x_i)[\tilde{T}]'\beta)^2] + \sigma^2 \frac{k}{n} \tag{1.6}$$

όπου $k = rank(E_n\,[x_i[\tilde{T}]x_i[\tilde{T}]'])$). Το μοντέλο *oracle* γνωρίζει το κίνδυνο για κάθε ένα από τα μοντέλα $\tilde{T}$ και μπορεί να ελαχιστοποιήσει αυτό τον κίνδυνο

$$min_{\tilde{T}} min_{\beta \in R^{|\tilde{T}|}} E_n[(f_i - x_i)[\tilde{T}]'\beta)^2] + \sigma^2 \frac{k}{n} \tag{1.7}$$

επιλέγοντας το καλύτερο μοντέλο ή το *oracle* μοντέλο T. Αυτό το πρόβλημα είναι στην πραγματικότητα ισοδύναμο με το (1.5), υπό την προϋπόθεση ότι $rank(E_n\,[x_i[\tilde{T}]x_i[\tilde{T}]']) = \|\beta_0\|_0$, δηλαδή να έχει πλήρη βαθμό. Επομένως, σε αυτή την περίπτωση η τιμή β_0 που λύνει το πρόβλημα (1.5) είναι η αναμενόμενη τιμή του *oracle* εκτιμητή των ελάχιστων τετραγώνων

$$\hat{\beta}_T = E_n[x_i[T]x_i[T]']^{-1}E_n[x_i[T]y_i] \tag{1.8}$$

, δηλαδή να ισχύει ότι $\beta_0 = E_n[x_i[T]x_i[T]']^{-1}E_n[x_i[T]f_i]$. Αυτή η τιμή είναι η πραγματική παράμετρος (*true parameter*) και το *oracle* μοντέλο T είναι το πραγματικό μοντέλο. Να τονισθεί ότι στην περίπτωση που έχουμε $c_s = 0$ τότε θα προκύπτει $f_i = x_i'\beta_0$, το οποίο μας δίνει μια ειδική παραμετρική περίπτωση.

1.3 Παραδείγματα

Ακολουθούν κάποια παραδείγματα για να μας δείξουν κατά πόσο εφαρμόζονται προσεγγιστικά ή όχι τα αραιά μοντέλα υψηλών διαστάσεων στην οικονομετρία.

1.3.1 Αραιά Μοντέλα για παλινδρομήσεις κέρδους

Στο συγκεκριμένο μοντέλο έχουμε μεταβλητή απόκρισης y_i ,η οποία αντιστοιχεί στο λογαριθμικό-μισθό, που είναι δεδομένης της μεταβλητής z_i η οποία αντιστοιχεί στην εκπαίδευση (μετριέται σε χρόνια φοίτησης). Με βάση μετρήσεις που αφορούν την εκπαίδευση που αποκτήθηκε σε συγκεκριμένο χρονικό διάστημα, εξάγουμε την σχέση του αναμενόμενου μισθού y_i δεδομένης της εκπαίδευσης z_i:

$$E[y_i|z_i] = \sum_{j=1}^{p} \beta_{0j} P_j(z_i) \qquad (1.9)$$

Στο πιο πάνω τύπο έγινε χρήση προσεγγιστικών συναρτήσεων $P_1(z_i), \dots, P_p(z_i)$ που μπορεί να είναι μετασχηματισμοί πολυωνύμων ή *splines* (κατά τμήματα πολυώνυμα) των z_i.

Δεδομένου ότι η παραπάνω συνάρτηση πιθανό να μην είναι μοναδική, μια συμβατική αραιή προσέγγιση που χρησιμοποιείται συνήθως στην οικονομετρία είναι η ακόλουθη:

$$f(z_i) := E[y_i|z_i] = \tilde{\beta}_1 P_1(z_i) + \cdots + \tilde{\beta}_s P_s(z_i) + \tilde{r}_i \qquad (1.10)$$

όπου P_j είναι είτε πολυώνυμα είτε κατά τμήματα πολυώνυμα χαμηλής τάξης με συνήθως $s = 4$ (Σχήμα 1.1) ή $s = 5$ (Σχήμα 1.2) όρους. Επίσης σε αυτή την περίπτωση δεν υπάρχει κάποια εγγύηση που να μας εξασφαλίζει ότι η προσέγγιση του σφάλματος $\tilde{r}_i$ θα είναι μικρή ή ότι τα πολυώνυμα που επιλέχθηκαν είναι τα καλύτερα προσεγγιστικά πολυώνυμα s-διαστάσεων .

Μέσα από την συνάρτηση που περιεγράφηκε αναμένουμε να παρατηρήσουμε την συμπεριφορά που παρουσιάζεται παραδείγματος χάριν σε υψηλά επίπεδα εκπαίδευσης (π.χ MBA). Παρόλα αυτά, το γεγονός ότι τα πολυώνυμα που επιλέγηκαν είναι χαμηλής τάξεως πιθανό να μην είναι σε θέση να συλλάβουν αυτή τη συμπεριφορά πολύ καλά και να έχουμε μεγάλο σφάλμα $\tilde{r}_i$. Για τον λόγο αυτό θα πρέπει να ψάξουμε και κάποια πολυώνυμα υψηλότερης τάξεως έτσι ώστε

να μας δώσουν μια καλύτερη προσέγγιση και μικρότερα σφάλματα. Το ζήτημα είναι κατά πόσο εύκολο είναι να βρούμε μια συνάρτηση του τύπου :

$$f(z_i) := E[y_i|z_i] = \beta_{k_1} P_{k_1}(z_i) + \cdots + \beta_{k_s} P_{k_s}(z_i) + r_i \qquad (1.11)$$

όπου οι δείκτες των ανεξάρτητων μεταβλητών $k_1, \dots, k_s$ επιλέγονται από το σύνολο $\{1, \dots p\}$ και να μας δίνει καλύτερα αποτελέσματα από την συνάρτηση (1.10).

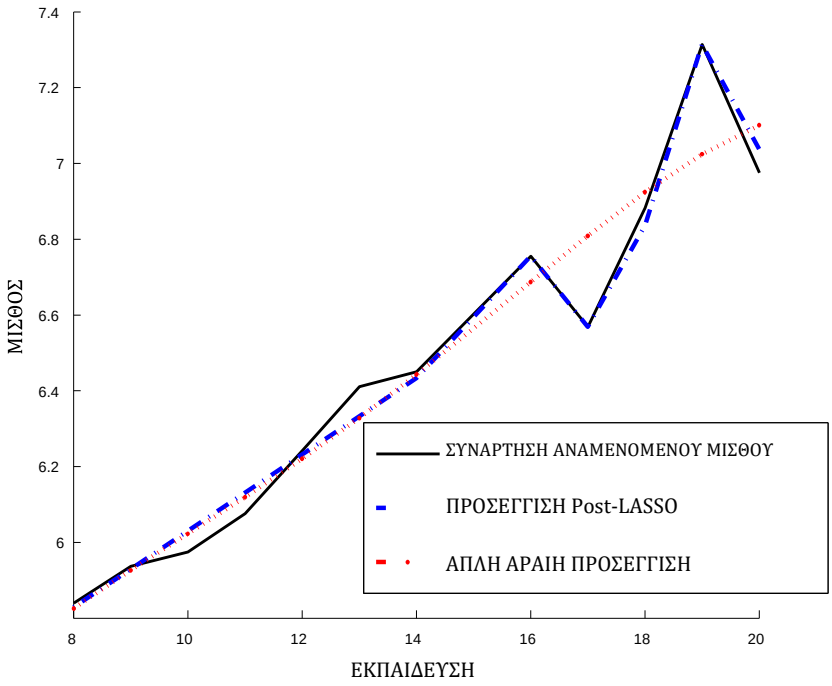
Η κατασκευή αυτής της προσεγγιστικά καλύτερης συνάρτησης εξαρτάται από το κατά πόσο είναι σύνθετη ή απλή η συμπεριφορά της πραγματικής συνάρτησης (1.9). Στην περίπτωση που είναι απλά τα πράγματα τότε προφανώς η προσέγγιση (1.11) είναι καλύτερη της (1.10) αφού θα υπάρχουν κάποιοι σημαντικοί όροι υψηλότερης τάξης. Με τον τρόπο αυτό η συνάρτηση πλέον θα μπορούσε να παρουσιάσει θετικές αλλαγές και να είναι σε θέση να δώσει καλύτερη συμπεριφορά για υψηλά επίπεδα εκπαίδευσης όπως είναι το MBA.

Όλο αυτό πάμε να το δούμε αναλυτικότερα σε μια εφαρμογή για $s = 4$ όρους. Συγκεκριμένα, ας εξετάσουμε το εισόδημα νεαρών λευκών ανδρών για το έτος 2000 (Angrist, Chernozhukov και Fernandez-Val (2006)). Θεωρούμε ότι έχουμε να κάνουμε με δεδομένα πληθυσμού και κάνουμε χρήση της συνάρτησης $f(z_i) := E[y_i|z_i]$ χωρίς σφάλματα. Έτσι σχεδιάζουμε την αναμενόμενη συνάρτηση μισθού (μαύρο χρώμα) στο Σχήμα 1.1 . Στο ίδιο Σχήμα σχεδιάζουμε μια αραιή προσέγγιση (κόκκινο χρώμα) που είναι της μορφής (1.10) με τα $P_1, \dots, P_s$ που αντιπροσωπεύει πολυώνυμο $(s-1)$-βαθμού. Επιπλέον, στο ίδιο Σχήμα σχεδιάζουμε ακόμη μια αραιή προσέγγιση, της μορφής ((1.11), μπλε χρώμα) με τα $P_{k_1}, \dots, P_{k_s}$ που αποτελείται από ένα σταθερό, ένα γραμμικό όρο και δύο γραμμικούς όρους *splines* με κόμβους που βρίσκονται στα 16 και 19 χρόνια της εκπαίδευσης (στην περίπτωση με $s = 5$ υπάρχει και ένας τρίτος κόμβος στα 17).

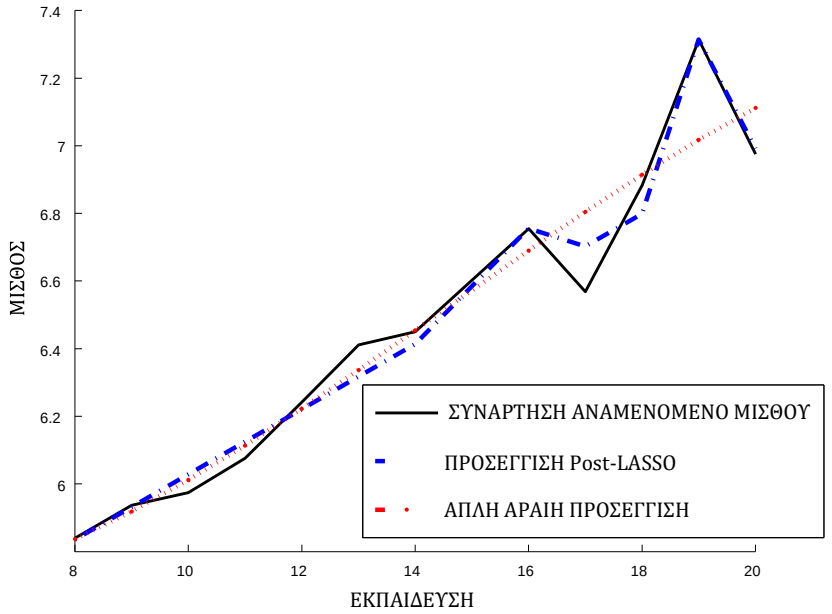
Όσον αφορά την δεύτερη προσέγγιση που είναι της μορφής (1.11) μπορούμε να την βρούμε επίσης χρησιμοποιώντας τις l_1-ποινικοποιημένες μεθόδους, αν και σε αυτή την περίπτωση θα μπορούσαμε να κατασκευάσουμε μια προσέγγιση βλέποντας μόνο το Σχήμα 1.1 αφού το μεγαλύτερο μέρος της συνάρτησης

περιγράφεται από μια γραμμική με μερικές απότομες αλλαγές, στα σημεία που τονίσθηκε πριν η ύπαρξη των κόμβων, οι οποίες μπορούν να αντιπροσωπευθούν από γραμμικούς spline όρους.

Τέλος, όσον αφορά προσεγγίσεις χαμηλών διαστάσεων, απαιτείται να γίνει η έρευνα σε ένα πολύ μεγάλο σύνολο μοντέλων. Για τον λόγο αυτό χρησιμοποιήσαμε l_1-ποινικοποίηση ελάχιστων τετραγώνων η οποία ποινικοποιεί το μέγεθος των μοντέλων μέσα από το άθροισμα των απόλυτων τιμών των ανεξάρτητων μεταβλητών.



Σχήμα 1.1: Αναπαριστά την αραιή προσέγγιση κατά Post-LASSO και την παραδοσιακή (πολυώνυμο χαμηλού βαθμού) προσέγγιση της συνάρτησης του μισθού με $s = 4$.



Σχήμα 1.2: Αναπαριστά την αραιή προσέγγιση κατά Post-LASSO και την παραδοσιακή (πολυώνυμο χαμηλού βαθμού) προσέγγιση της συνάρτησης του μισθού με $s = 5$.

1.3.2 Εκτιμήσεις σειρών και η συνθήκη ASM

Όπως αναφέρθηκε και πιο πάνω υπάρχει μια συσχέτιση μεταξύ της συνθήκης ASM και με τις προσεγγίσεις των σειρών των συναρτήσεων παλινδρόμησης του Newey (1997) με κάποιες διαφορές όμως.

Αναλυτικότερα, έστω ότι έχουμε το σύνολο $\{P_j(z), j \geq 1\}$ συναρτήσεων ορθοκανονικής βάσης στο $[0,1]^d$. Ένα παράδειγμα μπορεί να είναι τα ορθοπολυώνυμα με βάση το μέτρο Lebesgue [3]. Υποθέτουμε για λόγους απλότητας ότι τα z_i έχουν ομοιόμορφη κατανομή στο $[0,1]^d$.

Αφού υποθέσουμε ότι $E[f^2(z_i)] < \infty$, αναπαριστούμε την συνάρτηση μέσω μιας επέκτασης Fourier, δηλαδή $(z) = \sum_{j=1}^{\infty} \delta_j P_j(z)$, όπου $\{\delta_j, j \geq 1\}$ είναι οι συντελεστές Fourier οι οποίοι ικανοποιούν την σχέση $\sum_{j=1}^{\infty} {\delta_j}^2 < \infty$. Επιπλέον, θεωρούμε ότι η συνάρτηση f είναι ομαλή γεγονός που σημαίνει ότι οι συντελεστές Fourier, δ_j, προβάλλουν ένα πολυώνυμο decay $\delta_j \propto j^{-\nu}$, όπου το ν είναι ένα μέτρο ομαλότητας της συνάρτησης f.

Σε αντίθεση με την συνθήκη *ASM*, εμείς τώρα εξετάζουμε την επέκταση της σειράς που χρησιμοποιεί τους K πρώτους όρους για προσέγγιση και άρα:

$$f(z) = \sum_{j=1}^{K} \beta_{0j} P_j(z) + \alpha_c(z), \text{ με } \beta_{0j} = \delta_j \qquad (1.12)$$

Ο όρος $\alpha_c(z)$ αντιστοιχεί στο σφάλμα προσέγγισης το οποίο ικανοποιεί την ακόλουθη ανισότητα:

$$\sqrt{E_n[{\alpha_c}^2(z)]} \lesssim_P \sqrt{E[{\alpha_c}^2(z)]} \lesssim K^{\frac{-2\nu+1}{2}} \qquad (1.13)$$

Εξισορροπώντας τον όρο $K^{\frac{-2\nu+1}{2}}$ του σφάλματος προσέγγισης με τον όρο $\sqrt{K/n}$ του σφάλματος εκτίμησης παίρνουμε τον αριθμό του oracle-ρυθμού-

[3] Το μέτρο Lebesgue ενός συνόλου είναι το κάτω πέρας του αθροίσματος των μηκών των ημιάνοικτων διαστημάτων, η ένωση των οποίων αποτελεί κάλυψη για το σύνολο. Για $\bar{L}: P(R) \to [0, \infty]$ ορίζεται ως το $\bar{L}(A) = \inf\{\sum_{i \in N}(b_i - a_i)$ $\acute{\omega}\sigma\tau\varepsilon$ $A \subset \cup_{i \in N}[a_i, b_i)$ $\mu\varepsilon$ $\alpha_i, b_i \in R$ $\kappa\alpha\iota$ $\alpha_i < b_i\}$

βελτιστοποίησης των όρων της σειράς $s = K \propto n^{\frac{1}{2\nu}}$, καθώς και ο εκτιμητής oracle, ο οποίος γνωρίζει το s, θα εκτιμήσει την συνάρτηση μας στο ρυθμό oracle του $n^{\frac{1-2\nu}{4\nu}}$.

Με τον τρόπο αυτό θα γνωρίζουμε και την ταυτότητα των σημαντικότερων όρων της σειράς $T = \{1, \dots, s\}$ οι οποίοι θα είναι οι πρώτοι s όροι. Έτσι, συμπεραίνουμε ότι για την συνθήκη *ASM* θα έχουμε ότι $f(z) = \sum_{j=1}^{P} \beta_{0j} P_j(z) + a(z)$ με $\beta_{0j} = \delta_j$ για $j \leq s$ και $\beta_{0j} = 0$ για $s + 1 \leq j \leq p$ και $a(z_i) = \alpha_c(z_i)$, το οποίο συμπίπτει με την σειρά του Newey έτσι ώστε $\sqrt{E_n[\alpha^2(z_i)]} \lesssim_P \sqrt{s/n}$ και $\|\beta_{0j}\|_0 \leq s$.

Ακολούθως υποθέτουμε ότι για τους συντελεστές Fourier δ_j ισχύουν τα πιο κάτω:

$$\delta_j = 0 \text{ για } j \leq M \text{ και } \delta_j \propto (j - M)^{-\nu} \text{ για } j > M \qquad (1.14)$$

Το σίγουρο είναι ότι εδώ η προσέγγιση της σειράς θα βασιστεί στους πρώτους $K \leq M$ όρους και η σειρά $\sum_{j=1}^{K} \delta_j f_j(z)$ δεν θα έχει καμία δυνατότητα πρόβλεψης και ο αντίστοιχος εκτιμητής της σειράς που βασίζεται στους πρώτους K όρους θα αποτύχει εντελώς.

Από την άλλη, με χρήση της συνθήκης ASM οι εκτιμητές θα έχουν απόδοση που θα είναι κοντά επίπεδο oracle. Έτσι, χρησιμοποιώντας τους πρώτους p όρους της σειράς σχηματίζουμε την ακόλουθη προσέγγιση της σειράς:

$$f(z) = \sum_{j=1}^{P} \beta_{0j} P_j(z) + a(z) \qquad (1.15)$$

όπου $\beta_{0j} = 0$ για $j \leq M$ και $j > M + s$, $\beta_{0j} = \delta_j$ για $M + 1 \leq j \leq M + s$ με $s \propto n^{\frac{1}{2\nu}}$, και το p τέτοιο ώστε $M + n^{\frac{1}{2\nu}} = o(p)$.

Επομένως, $\|\beta_0\|_0 = s$, και έχουμε ότι

$$\sqrt{E_n[\alpha^2(z_i)]} \lesssim_P \sqrt{E[\alpha^2(z_i)]} \lesssim \sqrt{\frac{s}{n}} \lesssim n^{\frac{1-2\nu}{4\nu}} \qquad (1.16)$$

ΚΕΦΑΛΑΙΟ 2
ΑΡΑΙΕΣ ΜΕΘΟΔΟΙ ΕΚΤΙΜΗΣΗΣ
(ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΕΣ ΜΕΘΟΔΟΙ LASSO)

2.1 Ποινικοποιημένες μεθόδοι

Στον τομέα της στατιστικής έχουν αναπτυχθεί διάφορες μέθοδοι για την επιλογή μεταβλητών. Τις χρησιμοποιούμε σε μοντέλα με αρκετές ανεξάρτητες μεταβλητές με σκοπό να επιλέξουμε τις στατιστικά σημαντικές μεταβλητές, αυτές δηλαδή που επηρεάζουν σημαντικά την μεταβλητή απόκρισης y. Οι πιο γνωστές και συχνότερα χρησιμοποιούμενες είναι η κατά βήματα απαλοιφή (*stepwise*) και η μέθοδος επιλογής καλύτερου υποσυνόλου (*best subset selection*). Παρόλα αυτά, χαρακτηρίζονται από υψηλή μεταβλητότητα και χαμηλή ακρίβεια πρόβλεψης· ειδικά όταν ο αριθμός των μεταβλητών πρόβλεψης είναι πολύ μεγάλος.

Κατά την τελευταία δεκαετία, έγινε έντονη η παρουσία των ποινικοποιημένων μεθόδων παλινδρόμησης (*penalized regression methods*) οι οποίες παρέχουν καλύτερα αποτελέσματα έναντι των παραδοσιακών μεθόδων επιλογής. Συνάμα, αποτελούν μια προσέγγιση για να αποφευχθεί η μη αντιστρεψιμότητα του $(X^T X)$ που παρουσιάζεται όταν $p > n$.

2.1.1 Ποινικοποιημένα ελάχιστα τετράγωνα

Προτού καταλήξουμε στις διάφορες μορφές των εκτιμητών LASSO, θα αναφερθούμε στα Ποινικοποιημένα ελάχιστα τετράγωνα (*Penalized least squares, PLS*).

Υποθέτουμε ότι έχουμε τα δεδομένα της μορφής $(x_i^T, y_i)_{i=1}^n$ όπου y_i είναι η i-οστή παρατήρηση της μεταβλητής απόκρισης (*response variable*) και x_i είναι

το συσχετισμένο διάνυσμα p-διαστάσεων των επεξηγηματικών μεταβλητών. Υποθέτουμε ότι τα δεδομένα είναι ένα τυχαίο δείγμα από τον πληθυσμό (x^T, y) και η μεταβλητή απόκρισης y έχει μέσο που εξαρτάται από ένα γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών $\beta^T x$ όπου $\beta = (\beta_1, \dots, \beta_p)^T$.

Στα μοντέλα υψηλών διαστάσεων, υποθέτουμε ιδανικά ότι οι περισσότεροι παράμετροι β_j είναι ακριβώς μηδέν, γεγονός που σημαίνει ότι μόνο μερικές από τις ανεξάρτητες μεταβλητές συμβάλλουν στην μεταβλητή απόκρισης. Ο στόχος της επιλογής μεταβλητής, είναι να προσδιορίζει τις πολύ σημαντικές ανεξάρτητες μεταβλητές οι οποίες θα έχουν συντελεστές διάφορους του μηδενός και να δίνει τις ακριβείς εκτιμήσεις αυτών των παραμέτρων.

2.1.1.1 PLS μια μεταβλητής

Έστω ότι έχουμε το γραμμικό μοντέλο παλινδρόμησης

$$y = X\beta + \varepsilon \tag{2.1}$$

όπου $y = (y_1, \dots, y_n)^T$ είναι ένα διάνυσμα απόκρισης n-διαστάσεων , $X = (x_1, \dots, x_n)^T$ είναι ένας $n \times p$ πίνακας σχεδιασμού και ε είναι ένα n-διαστάσεων διάνυσμα θορύβου. Εξετάζουμε την ειδική περίπτωση του κανονικού γραμμικού μοντέλου με τον ορθοκανονικό πίνακα σχεδιασμού , δηλαδή $X^T X = nI_p$.

Το πρόβλημα *PLS* θα είναι :

$$min_{\beta \in R^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \right\} \tag{2.2}$$

όπου $\|\cdot\|_2$ ορίζει την L_2 νόρμα και $p_\lambda(\cdot)$ είναι η συνάρτηση ποινής με δείκτη την κανονικοποιημένη παράμετρο $\lambda \geq 0$. Με την κανονικοποίηση της εκτίμησης ελάχιστων τετραγώνων, επιθυμούμε ταυτόχρονα να επιλέξουμε τις σημαντικές

μεταβλητές και να εκτιμήσουμε τους παράγοντες παλινδρόμησης τους με αραιές εκτιμήσεις.

2.1.1.2 PLS πολλών μεταβλητών

Εξετάζουμε το πρόβλημα πολλών μεταβλητών των ποινικοποιημένων ελάχιστων τετραγώνων (2.2) με το γενικό πίνακα σχεδιασμού X. Ο στόχος είναι να εκτιμήσουμε τους παράγοντες του πραγματικού αραιού διανύσματος παλινδρόμησης $\beta_0 = (\beta_{0,1}, \dots, \beta_{0,p})^T$ στο γραμμικό μοντέλο (2.1) όπου ο αριθμός των παραμέτρων είναι πολύ μεγαλύτερος από το μέγεθος του δείγματος · $p > n$.

Η l_0-κανονικοποίηση εμφανίζεται σε πολλές κλασσικές μεθόδους επιλογής μοντέλου όπως η AIC (*Akaike Information Criterion*) του Akaike (1973,1974) και η BIC (*Bayesian Information Criterion*) του Schwartz (1978). Ως γνωστόν η $l_0 - \nu\acute{o}\rho\mu\alpha$ είναι το άθροισμα των μη μηδενικών συντελεστών, δηλαδή $\|\beta\|_0 = \sum_{j=1}^{p} 1\{|\beta_j| > 0\}$. Ισοδυναμεί με την καλύτερη επιλογή υποσυνόλου και έχει αποδειχθεί ότι έχει πολύ καλές ιδιότητες δειγματοληψίας. Εντούτοις η χρήση των εκτιμητών AIC και BIC καθίσταται πρακτικά αδύνατη, μιας και για την επίλυση του προβλήματος πιθανό να απαιτηθούν οι λύσεις $\sum_{\kappa \le n} \binom{p}{k}$ προβλημάτων ελάχιστων τετραγώνων.

Είναι προφανές ότι η αυξημένη υπολογιστική πολυπλοκότητα της επίλυσης του πιο πάνω προβλήματος καθιστά επιτακτική την εύρεση εναλλακτικών μεθόδων για την αναζήτηση αραιών λύσεων. Συγκεκριμένα αυτές οι υπολογιστικές δυσκολίες οδήγησαν σε διάφορες συνεχείς χαλαρώσεις της ασυνεχούς ποινής l_0. Παραδείγματος χάριν, η παλινδρόμηση bridge των Frank και Friedman (1993) χρησιμοποιεί την ποινή L_q όπου $0 < q < 2$. Ειδικότερα, η χρήση της ποινής l_2 ονομάζεται ridge παλινδρόμηση. Ο μη αρνητικός βρόγχος (*non-negative garotte*) εισήχθη από τον Breiman (1995) για την επιλογή του μοντέλου και την συρρίκνωση της εκτίμησης. Η l_1- ποινικοποιημένη μέθοδος ελάχιστων τετραγώνων (l_1 *penalized least squares method*) ονομάστηκε LASSO (*Least Absolute Shrinkage and Selection Operator*) από τον Tibshirani (1996).

Ανάμεσα στις συχνά χρησιμοποιημένες συναρτήσεις ποινής, συμπεριλαμβάνονται οι SCAD των Fan και Li (2001) και η MCP του Zhang (2010). Μια οικογένεια των κοίλων ποινικοποιήσεων που γεφυρώνει τις ποινικοποιήσεις l_0 και l_1 εισάχθηκε από τους Lv και Fan (2009) με σκοπό την επιλογή μοντέλου και την αραιή ανάκτηση. Ένας γραμμικός συνδυασμός των l_1 και l_2 ποινικοποιήσεων ονομάστηκε ελαστικό δίκτυο (*elastic net*) από τους Zou & Hastie (2005).

Παρόλα αυτά εμείς θα ασχοληθούμε σε αυτό το κεφάλαιο με την μέθοδο παλινδρόμησης με βάση τον εκτιμητή LASSO. Ακολουθεί μια σύντομη περιγραφική αναφορά για τα διάφορα είδη των εκτιμητών LASSO, ενώ πιο αναλυτικά αποτελέσματα σχετικά με αυτούς θα δοθούν σε επόμενα κεφάλαια.

2.2 Περιγραφή των διάφορων παραλλαγών της μεθόδου LASSO

2.2.1 Εκτιμητής LASSO

Με την χρήση της ποινής l_1 πλέον η εκτίμηση θα γίνεται από τον εκτιμητή LASSO $\hat{\beta}$ ο οποίος ελαχιστοποιεί μια κυρτή συνάρτηση και αποτελεί λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$min_{\beta \in R^{\rho}} E_n[(y_i - x_i'\beta)^2] + \frac{\lambda}{n}\|\beta\|_1 \qquad (2.3)$$

όπου $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$, είναι δηλαδή το άθροισμα των απόλυτων τιμών των στοιχείων του διανύσματος s.

Σχετικά με την επιλογή του επίπεδου ποινής λ (*penalty level*), προτάθηκε το 2009 από τους Bickel, Ritov και Tsybakov η ακόλουθη σχέση με το σκεπτικό ότι θα δίνει κοντινές oracle τιμές του ρυθμού σύγκλισης του εκτιμητή :

$$\lambda = 2 \cdot c\sigma\sqrt{2nlog(2p/\gamma)} \qquad (2.4)$$

όπου $c > 1$ και $1 - \gamma$ είναι το επίπεδο εμπιστοσύνης που θα πρέπει να είναι όσο πιο κοντά γίνεται στη μονάδα.

Εντούτοις, παρουσιάζεται μια υπολογιστική δυσκολία μιας και το συγκεκριμένο επίπεδο ποινής όπως φαίνεται και από τον ορισμό του εξαρτάται από το άγνωστο σ. Για τον λόγο αυτό οι Belloni και Chernozhukov το 2011 όρισαν ένα διαφορετικό τύπο για το επίπεδο ποινής.

2.2.1.1 Score S - Αποτελεσματικός θόρυβος

Προτού παρουσιάσουμε τα αποτελέσματα τους σχετικά με το επίπεδο ποινής, να ορίσουμε την ποσότητα S που ονομάζεται «*score*» και αποτελεί τον αποτελεσματικό θόρυβο (*noise*) του προβλήματος. Το *score* ορίζεται ως η κλίση της συνάρτησης $\hat{Q}$ στην πραγματική της τιμή β_0, δηλαδή $S = \forall \hat{Q}(\beta_0)$ όπου $\hat{Q} = E_n[(y_i - x_i'\beta)^2]$ και με αντικατάσταση θα προκύψει ότι $S = 2E_n[x_i\varepsilon_i]$. Η χρήση του S είναι ότι συμβάλλει στην επιλογή του επίπεδου ποινής μιας και επιθυμούμε να επιλέξουμε το μικρότερο επίπεδο ποινής έτσι ώστε να ισχύει η ακόλουθη ανισότητα:

$$\lambda > c\Lambda \,, για\ \Lambda \coloneqq n\|S\|_\infty \ \Rightarrow\ \lambda \geq cn\|S\|_\infty \qquad (2.5)$$

με πιθανότητα τουλάχιστον $1 - \gamma$, όπου το $1 - \gamma$ θα πρέπει να είναι όσο το δυνατό πιο κοντά γίνεται στη μονάδα και όπου Λ είναι το μέγιστο αποτέλεσμα που εξαρτάται από το n ενώ το $c > 1$ είναι η θεωρητική σταθερά των Bicker, Ritor και Tsybakov (2009).

Επίσης, ισχύει ότι $\|S/(2\sigma)\|_\infty =_d max_{1 \leq j \leq p}|E_n[x_i g_i]|$ όπου τα g_i είναι ανεξάρτητες μεταβλητές ταυτόσημα κατανεμημένες στη $N(0,1)$.

Με χρήση των παραπάνω σχέσεων οι Belloni και Chernozhukov εξήγαγαν τους παρακάτω τύπους:

X-ανεξάρτητο επίπεδο ποινής: $\lambda = 2 \cdot c\hat{\sigma}\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right)$ (2.6)

όπου $c > 1$ και $1 - \gamma$ είναι το επίπεδο εμπιστοσύνης .

X-εξαρτημένο επίπεδο ποινής: $\lambda = c \cdot 2\hat{\sigma}\Lambda(1-\gamma|X)$ (2.7)

όπου $\Lambda(1-\gamma|X) = (1-\gamma) -$ το ποσοστιαίο σημείο του $n\|E_n[x_i g_i]\|_\infty|X$,

$X = [x_1, x_2, \dots, x_n]'$ και τα g_i είναι ανεξάρτητες μεταβλητές ταυτόσημα κατανεμημένες με $N(0,1)$, οι οποίες μπορούν εύκολα να προσεγγιστούν με προσομοίωση (μοντελοποίηση).

Γενικά, προτιμάται το X-εξαρτημένο επίπεδο ποινής, δεδομένου ότι εκ κατασκευής προσαρμόζεται στον πίνακα σχεδιασμού X και είναι λιγότερο συντηρητικός από τον άλλο λαμβάνοντας υπόψη και την ακόλουθη σχέση που παρεμπιπτόντως μας δίνει και ένα ανώτατο όριο για το επίπεδο ποινής, λ:

$$\Lambda(1-\gamma|X) \leq \sqrt{n}\Phi^{-1}\left(1-\frac{\gamma}{2p}\right) \leq \sqrt{2nlog\left(\frac{2p}{\gamma}\right)} \quad (2.8)$$

2.2.1.2 Εκτιμητής θορύβου $\hat{\sigma}$

Ο εκτιμητής του θορύβου (*noise*) $\hat{\sigma}$ που χρησιμοποιείται στους παραπάνω τύπους δίνεται από τη σχέση $\hat{\sigma} = \sigma + o_P(1)$ που υπολογίζεται μέσω μιας επαναληπτικής μεθόδου.

Πρώτα ορίζουμε την αρχική εκτίμηση ως $\hat{\sigma}^0 = \sqrt{Var_n(y_i)} \coloneqq \sqrt{E_n[(y_i - \bar{y})^2]}$ όπου $\bar{y} = E_n[y_i]$. Ο $\hat{\sigma}^0$ είναι συντηρητικός εκτιμητής με $\hat{\sigma}^0 = \sigma^0 + o_P(1)$ όπου $\sigma^0 = \sqrt{Var(y_i)} \geq \sigma$ δεδομένου ότι η x_i περιλαμβάνει μια σταθερά. Ακολούθως ορίζουμε την βελτιωμένη εκτίμηση $\hat{\sigma} = \sqrt{\hat{Q}(\hat{\beta})}$. Έπειτα, χρησιμοποιούμε την εκλεπτυσμένη εκτίμηση (*refined estimate*) $\hat{\sigma}^2$ για να πάρουμε τον εκλεπτυσμένο εκτιμητή LASSO $\hat{\beta}$. Μπορούμε να σταματήσουμε εδώ ή να επαναλάβουμε τα δύο τελευταία βήματα.

Με λίγα λόγια ο αλγόριθμος είναι:

Ορίζουμε $\hat{\sigma}^0 = \sqrt{Var_n(y_i)}$ και $k = 0$ όπως επίσης καθορίζουμε μια μικρή σταθερά $\nu > 0$, το επίπεδο ανοχής , και την σταθερά $I > 1$,το άνω φράγμα του αριθμού των επαναλήψεων.

(1) Υπολογίζουμε τον εκτιμητή LASSO $\hat{\beta}$ με βάση το επίπεδο ποινής $\lambda = c \cdot 2\,\hat{\sigma}^k \Lambda(1-\gamma|X)$.

(2) Ορίζουμε $\hat{\sigma}^{k+1} = \sqrt{\hat{Q}(\hat{\beta})}$.

(3) Εάν $|\hat{\sigma}^{k+1} - \hat{\sigma}^k| \leq \nu$ ή $k+1 \geq I$ τότε σταματάμε και $\hat{\sigma} = \hat{\sigma}^{k+1}$, αλλιώς $k \leftarrow k+1$ και επαναλαμβάνουμε το βήμα **(1)** .

2.2.2 Iterated LASSO

Μαζί με όλα τ' άλλα, το 2010 οι Belloni, Chen, Chernozhukov και Hansen διατύπωσαν μια εφικτή διαδικασία LASSO, την επονομαζόμενη Επαναλαμβανόμενη LASSO (*Iterated LASSO*), για τις περιπτώσεις με ετεροσκεδαστικότητα, των μη Γκαουσιανών διαταραχών.

Ο εκτιμητής της Επαναλαμβανόμενης LASSO θα έχει παρόμοια στατιστικά αποτελέσματα με αυτά του πιο πάνω μη εφικτού εκτιμητή και θα χρησιμοποιεί και αυτός με τη σειρά του τα επίπεδα ποινής που δίνονται από τις σχέσεις (2.6) και (2.7).

2.2.3 Square-root LASSO

Στη συνέχεια προτείνεται μια παραλλαγή από τους Belloni, Chernozhukov και Wang (2010) η οποία αν και έχει παρόμοια στατιστική απόδοση με τη LASSO, το επίπεδο ποινής της, λ, είναι ανεξάρτητο από την τιμή σ. Ορίζεται ως η Τετραγωνική-ρίζα LASSO (*Square-root LASSO*) εκτιμήτρια $\hat{\beta}$ και αποτελεί την λύση του ακόλουθου προβλήματος:

$$min_{\beta \in R^{\rho}} \sqrt{E_n[(y_i - x_i'\beta)^2]} + \frac{\lambda}{n}\|\beta\|_1 \qquad (2.9)$$

με επίπεδο ποινής $\lambda = c \cdot \tilde{\Lambda}(1-\gamma|X)$ (2.10)

όπου $c > 1$ και

$\tilde{\Lambda}(1-\gamma|X) = (1-\gamma) -$ το ποσοστιαίο σημείο του $n\|E_n[x_i g_i]\|_{\infty} / \sqrt{E_n[g_i{}^2]} \; |X$

με $g_i \sim N(0,1)$ για κάθε $i = 1,2,\dots,n$.

Στην περίπτωση της Τετραγωνικής-ρίζα LASSO εκτιμήτριας , η ασυμπτωτική επιλογή για το επίπεδο ποινής δίνεται από τον τύπο :

$$\lambda = c\Phi^{-1}\left(1 - \frac{\gamma}{2p}\right) \qquad (2.11)$$

Η Τετραγωνική-ρίζα LASSO εκτιμήτρια $\hat{\beta}$ αποτελεί επιπλέον και λύση του κωνικού προβλήματος προγραμματισμού:

$$min_{t \geq 0. \beta \in R^{\rho}} \; t + \frac{\lambda}{n}\|\beta\|_1 : \sqrt{E_n[(y_i - x_i'\beta)^2]} \leq t \qquad (2.12)$$

2.2.4 Post-LASSO

Γενικά, είναι γεγονός ότι η χρήση της l_1-νόρμα στην κανονικοποιήση βοηθά τους διάφορους εκτιμητές LASSO που παρουσιάστηκαν, για να αποφύγουν την υπερπροσαρμογή (*overfitting*) [4] των δεδομένων. Παρόλα αυτά για να επιτευχθεί αυτό συρρικνώνονται σταθεροί συντελεστές προς το μηδέν, προκαλώντας ενδεχομένως μια σημαντική μεροληψία. Για να αφαιρεθεί ένα μέρος αυτής της μεροληψίας, θεωρούμε τον εκτιμητή Post-LASSO $\tilde{\beta}$ ο οποίος εφαρμόζει την

[4] Η υπερπροσαρμογή συμβαίνει όταν ένα στατιστικό μοντέλο περιγράφει το τυχαίο σφάλμα ή το θόρυβο αντί της υποκείμενης σχέσης. Γενικά η υπερπροσαρμογή εμφανίζεται όταν ένα μοντέλο είναι υπερβολικά πολύπλοκο, όπως όταν έχει πάρα πολλές παραμέτρους σε σχέση με τον αριθμό των παρατηρήσεων. Ένα μοντέλο το οποίο έχει υπερπροσαρμογή θα έχει γενικά κακή προγνωστική απόδοση, καθώς μπορεί να διογκωθούν μικρές διακυμάνσεις στα δεδομένα

συνήθης παλινδρόμηση ελάχιστων τετραγώνων στο μοντέλο $\hat{T}$ το οποίο ως γνωστό επιλέγεται από τον l_1-κανονικοποιημένο εκτιμητή $\hat{\beta}$ (*LASSO*) .

Ορίζουμε τόσο για το μοντέλο όσο και για τον εκτιμητή Post-LASSO $\tilde{\beta}$ ότι :

$$\hat{T} = support(\hat{\beta}) = \{j \in \{1, \dots, p\} : |\hat{\beta}_j| > 0\}$$
$$\tilde{\beta} \in argmin_{\beta \in R^p} E_n[(y_i - x_i'\beta)^2] : \beta_j = 0 \text{ για κάθε } j \in \hat{T}^c \quad (2.13)$$

όπου $\hat{T}^c = \{1, \dots, p\} \setminus \hat{T}$.

Δηλαδή, ο εκτιμητής Post-LASSO $\tilde{\beta}$ είναι ο κλασσικός εκτιμητής ελάχιστων τετραγώνων που εφαρμόζεται στα δεδομένα μας μετά την αφαίρεση των συντελεστών παλινδρόμησης που δεν επιλέχθηκαν από το μοντέλο $\hat{T}$. Ομοίως υπάρχει και ο Post-Square-root LASSO. Στην περίπτωση που έχουμε την επιλογή του τέλειου μοντέλου, δηλαδή να ισχύει ότι $\hat{T} = T$ τότε αυτό σημαίνει ότι ο εκτιμητής Post-LASSO συμπίπτει με τον εκτιμητή oracle. Στην πραγματικότητα όμως κάτι τέτοιο είναι ανέφικτο αφού συνήθως θα προκύπτει ότι $\hat{T} \neq T$ και πιο συγκεκριμένα $\hat{T} \nsubseteq T$.

2.2.4.1 Εκτιμητής θορύβου $\hat{\sigma}$

Με βάση το ίδιο σκεπτικό που αναπτύχθηκε στην παράγραφο 2.2.1.2 ο αλγόριθμος για την εκτίμηση του θορύβου $\hat{\sigma}$ χρησιμοποιώντας επαναλήψεις του Post-LASSO θα είναι:

Ορίζουμε $\hat{\sigma}^0 = \sqrt{Var_n(y_i)}$ και $k = 0$ όπως επίσης καθορίζουμε μια μικρή σταθερά $\nu > 0$, το επίπεδο ανοχής , και την σταθερά $I > 1$,το άνω φράγμα του αριθμού των επαναλήψεων.

(1) Υπολογίζουμε τον εκτιμητή Post-LASSO $\tilde{\beta}$ με βάση το επίπεδο ποινής $\lambda = c \cdot 2\, \hat{\sigma}^k \Lambda(1 - \gamma|X)$.

(2) Ορίζουμε $\hat{\sigma}^{k+1} = \sqrt{\frac{n}{n-\hat{s}} \hat{Q}(\tilde{\beta})}$,όπου $\hat{s} = \|\tilde{\beta}\|_0 = |\hat{T}|$.

(3) Εάν $|\hat{\sigma}^{k+1} - \hat{\sigma}^{k}| \leq \nu$ ή $k+1 \geq I$ τότε σταματάμε και $\hat{\sigma} = \hat{\sigma}^{k+1}$, αλλιώς $k \leftarrow k+1$ και επαναλαμβάνουμε το βήμα **(1)** .

Μπορούμε να χρησιμοποιήσουμε $\lambda = c \cdot 2\,\hat{\sigma}^{k}\sqrt{n}\Phi^{-1}(1-\gamma/2\mathrm{p})$ στη θέση του X-εξαρτημένου επίπεδου ποινής. Να σημειωθεί σε αυτό το σημείο ότι με τη χρήση της LASSO για την εκτίμηση του σ (παράγραφος 2.2.1.2) προκύπτει ότι η ακολουθία $\hat{\sigma}^{k}$, $k \geq 2$ είναι μονότονη, ενώ χρησιμοποιώντας Post-LASSO για την εκτίμηση του $\hat{\sigma}^{k}$, $k \geq 1$ μπορούμε να υποθέσουμε ότι έχουμε ένα πεπερασμένο αριθμό διαφορετικών τιμών.

2.2.5 Adaptive LASSO

Εξαιτίας του γεγονότος ότι ο εκτιμητής LASSO δεν αποτελεί μια διαδικασία oracle θα παρουσιαστεί σε αυτή την παράγραφο μια παραλλαγή του που θα ικανοποιεί τις ιδιότητες oracle. Συγκεκριμένα μπορούμε να αντιστοιχίζουμε διαφορετικούς συντελεστές βαρύτητας (*weights*) σε διαφορετικούς συντελεστές β · πράγμα που γίνεται όταν θέλουμε να δώσουμε διαφορετική βαρύτητα (έμφαση) στις τιμές $x_1, \ldots, x_n$ ενός συνόλου δεδομένων. Ο σταθμισμένος εκτιμητής LASSO θα ορίζεται ως

$$argmin_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \qquad (2.14)$$

όπου w είναι το διάνυσμα των συντελεστών βαρύτητας (στάθμισης). Θα δειχθεί ότι αν οι συντελεστές βαρύτητας είναι εξαρτημένοι από τα δεδομένα και αν είναι σωστά επιλεγμένοι, τότε ο σταθμισμένος εκτιμητής LASSO θα ικανοποιεί τις ιδιότητες oracle. Η νέα αυτή μεθοδολογία ονομάζεται προσαρμοστική (*adaptive*) LASSO.

Επιλέγουμε $\gamma > 0$ και ορίζουμε το διάνυσμα των συντελεστών βαρύτητας να είναι $\widehat{w} = 1/|\hat{\beta}|^{\gamma}$. Τότε ο εκτιμητής adaptive LASSO $\hat{\beta}^{*(n)}$ θα ορίζεται ως εξής:

$$\hat{\beta}^{*(n)} = argmin_{\beta} \left\| y - \sum_{j=1}^{p} x_j \beta_j \right\|^2 + \lambda_n \sum_{j=1}^{p} w_j |\beta_j| \qquad (2.15)$$

Παρομοίως, μπορούμε να θέσουμε $A_n^* = \{j: \hat{\beta}_j^{*(n)} \neq 0\}$. Αξίζει να τονισθεί ότι το (2.15) είναι ένα κυρτό πρόβλημα βελτιστοποίησης, και ως εκ τούτου ο ελαχιστοποιητής του μπορεί να το λύσει αποτελεσματικά. Προφανώς και η adaptive LASSO είναι μια l_1 ποινικοποιημένη μέθοδος.

2.2.5.1 Oracle Ιδιότητες

Σε αυτή την παράγραφο θα δείξουμε ότι με μια σωστή επιλογή του επίπεδου ποινής λ_n, ο εκτιμητής adaptive LASSO μπορεί να χαρακτηρίζεται από τις ιδιότητες oracle.

Θεώρημα 2.1: Oracle ιδιότητες

Υποθέτουμε ότι $\lambda_n/\sqrt{n} \to 0$ και $\lambda_n n^{(\gamma-1)/2} \to \infty$. Τότε ο εκτιμητής adaptive LASSO θα πρέπει να ικανοποιεί τα ακόλουθα.

α. Συνέπεια [5] (*Consistency*) στην επιλογή μεταβλητής, δηλαδή $lim_n P(A_n^* = A) = 1$.

β. Ασυμπτωτική κανονικότητα [6] (*Asymptotic normality*), δηλαδή $\sqrt{n}(\hat{\beta}_A^{*(n)} - \beta_A^*) \xrightarrow{d} N(0, \sigma^2 \times C_{11}^{-1})$.

Παρατήρηση 2.1

Το Θεώρημα 2.1 δείχνει ότι αυτή η ποινή l_1 είναι τόσο καλή όσο οποιαδήποτε άλλη oracle ποινή. Τα δεδομένα που εξαρτώνται από το w είναι το κλειδί στο

[5] Γενικά μια ακολουθία τυχαίας μεταβλητής $\{\hat{\theta}_n\}$ καλείται συνεπής εκτιμητής της θ αν, $\hat{\theta}_n \xrightarrow{p} \theta$.

[6] Ένας συνεπής εκτιμητής $\hat{\theta}_n$ είναι ασυμπτωτικά κανονικός αν $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \Sigma)$. Ο πίνακας Σ ονομάζεται ασυμπτωτική διασπορά και συμβολίζεται ως $AVar(\hat{\theta}_n)$.

Θεώρημα 2.1. Καθώς το μέγεθος του δείγματος αυξάνεται, οι συντελεστές βαρύτητας για τους μηδενικούς συντελεστές πρόβλεψης διογκώνονται προς το άπειρο, ενώ οι συντελεστές βαρύτητας για τους μη μηδενικούς συντελεστές συγκλίνουν σε μια πεπερασμένη σταθερά. Έτσι μπορούμε ταυτόχρονα να εκτιμούμε αμερόληπτα μεγάλους συντελεστές και μικρά φράγματα εκτίμησης.

2.2.6 Γεωμετρική ερμηνεία των εκτιμητών LASSO και Post-LASSO

Θα ακολουθήσει μια γεωμετρική προσέγγιση σχετικά με τους εκτιμητές LASSO και Post-LASSO που περιεγράφηκαν πιο πάνω. Καταρχήν, διαπιστώνουμε ότι ο εκτιμητής LASSO, $\hat{\beta}$ μπορεί να λύνει και το ακόλουθο πρόγραμμα βελτιστοποίησης:

$$min_{\beta \in R^{\rho}} \|\beta\|_1 : \hat{Q}(\beta) \leq \gamma \qquad (2.16)$$

για κάποια τιμή $\gamma > 0$ η οποία εξαρτάται από το επίπεδο ποινής λ .

Αυτό που κάνει γεωμετρικά ο εκτιμητής LASSO είναι να αναζητά την ελάχιστη l_1-μπάλα που ονομάζεται διαμάντι (*diamond*) και έχει μη κενή τομή με ένα χαμηλότερο περίγραμμα του συνόλου της συνάρτησης κριτηρίου των ελάχιστων τετραγώνων, που ονομάζεται έλλειψη (*ellipse*). Παρακάτω αναπαρίστανται σε σχήματα το διαμάντι και η έλλειψη για 3 διαφορετικές περιπτώσεις και εξάγονται κάποια συμπεράσματα.

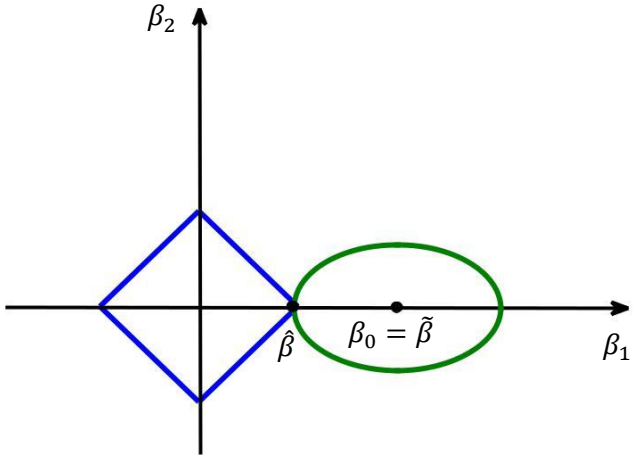
Στο Σχήμα 2.1, η έλλειψη αντιπροσωπεύει το χαμηλότερο περίγραμμα του συνόλου του πληθυσμού της συνάρτησης κριτηρίου $Q(\beta) = E[(y_i - x_i'\beta)^2]$ στην περίπτωση μηδενικού θορύβου ή στην περίπτωση άπειρου δείγματος. Στα Σχήματα 2.2 και 2.3 παρουσιάζεται το περίγραμμα συνόλου από το δείγμα της συνάρτησης κριτηρίου $\hat{Q}(\beta) = E_n[(y_i - x_i'\beta)^2]$ στο μη μηδενικό θόρυβο (μικρό και μεγάλο θόρυβο αντίστοιχα) ή στην περίπτωση του πεπερασμένου δείγματος. Επίσης, στο Σχήμα 2.2 απεικονίζεται η περίπτωση των δυο διαστάσεων

κατά την οποία οι συντεταγμένες της πραγματικής παραμέτρου β_0 να ισούνται με $(\beta_{01},\beta_{02}) = (1,0)$ και $T = support(\beta_0) = \{1\}$ με $s = 1$.

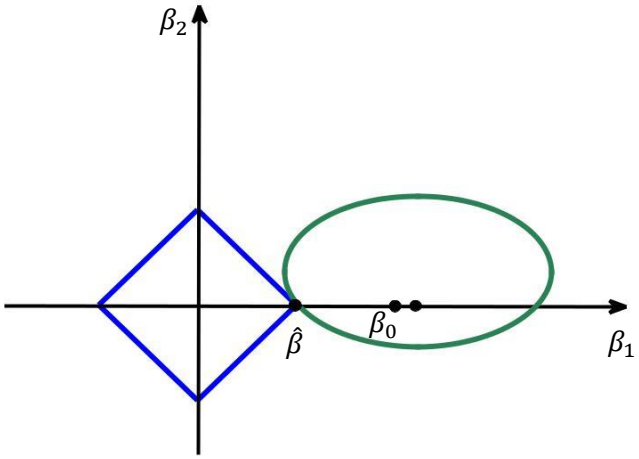
Η τομή των ελάχιστων διαμαντιών με τις ελλείψεις όπως παρουσιάζονται στα σχήματα μας δίνει το σύνολο των βέλτιστων λύσεων του εκτιμητή LASSO $\hat{\beta}$. Να σημειωθεί ότι στην πρώτη περίπτωση ο εκτιμητής της LASSO μπορεί εύκολα να αποκτήσει τη σωστή μορφή αραιότητας του β_0· αν και λόγω της κανονικοποίησης θα ωθείται μια μεγάλη μεροληψία προς το μηδέν. Από την άλλη, σχετικά με τον εκτιμητή Post-LASSO $\tilde{\beta}$ προκύπτει από το κέντρο της έλλειψης που τέμνεται από τον γραμμικό υπόχωρο που επιλέχθηκε από το LASSO. Με την σειρά του ο εκτιμητής Post-LASSO $\tilde{\beta}$ στην πρώτη περίπτωση αφαιρεί την μεροληψία και ανακτά εντελώς το β_0.

Στις περιπτώσεις 2.2 και 2.3 που ο θόρυβος είναι μη μηδενικός οι καμπύλες των συναρτήσεων κριτηρίων και τα κέντρα τους απομακρύνονται από τον αντίστοιχο πληθυσμό. Αναλυτικότερα, στην περίπτωση όπου έχουμε μικρό θόρυβο το εμπειρικό σφάλμα μετακινεί το κέντρο της έλλειψης σε ένα μη αραιό σημείο. Ωστόσο, ο εκτιμητής LASSO θέτει ορθά $\hat{\beta}_2 = 0$ και $\hat{\beta}_1 \neq 0$ αποκτώντας έτσι τη διάταξη αραιότητας του β_0. Με την σειρά του ο Post-LASSO $\tilde{\beta}$,με χρήση του φορέα γίνεται ο εκτιμητής oracle ο οποίος βελτιώνει δραστικά σε σχέση με τον LASSO. Ακολούθως, στην περίπτωση που έχουμε μεγάλο θόρυβο, τα μεγάλα εμπειρικά λάθη που εμφανίζονται ωθούν το κέντρο του χαμηλότερου περιγράμματος του συνόλου μακριά από τον αντίστοιχο πληθυσμό. Αυτά τα μεγάλα εμπειρικά σφάλματα κάνουν τον εκτιμητή LASSO να μην είναι αραιός, θέτοντας εσφαλμένα $\hat{\beta}_2 \neq 0$. Ως εκ τούτου ο εκτιμητής Post-LASSO δεν χρησιμοποιεί τον ακριβή φορέα $T = \{1\}$ αλλά χρησιμοποιεί $\hat{T} = \{1,2\}$. Σε αυτή την περίπτωση ο εκτιμητής Post-LASSO δεν συμπίπτει με τον εκτιμητή oracle. Και στις τρείς περιπτώσεις φαίνεται η συρρίκνωση της μεροληψίας προς το μηδέν στον εκτιμητή LASSO που γίνεται με τη χρήση της ποινής l_1-νόρμας. Τέλος, στις περιπτώσεις όπου ο εκτιμητής LASSO επιτυγχάνει μια καλή μορφή αραιότητας, ο εκτιμητής Post-LASSO βελτιώνει δραστικά σε σχέση με τον προηγούμενο.

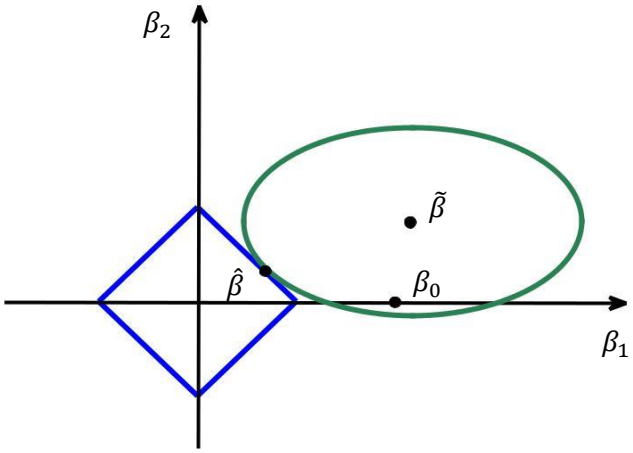
Σχήματα γεωμετρικής ερμηνείας



Σχήμα 2.1: Παρουσιάζει τη γεωμετρία των εκτιμητών LASSO και Post-LASSO χωρίς θόρυβο.



Σχήμα 2.2: Παρουσιάζει τη γεωμετρία των εκτιμητών LASSO και Post-LASSO με μικρό θόρυβο.



Σχήμα 2.3: Παρουσιάζει τη γεωμετρία των εκτιμητών LASSO και Post-LASSO με μεγάλο θόρυβο.

2.2.7 Εφαρμογή της Cross-Country Growth παλινδρόμησης

Σε αυτή την ενότητα θα εφαρμόσουμε τους εκτιμητές *LASSO* και *Post-LASSO* σε ένα παράδειγμα σχετικά με την διεθνή οικονομική ανάπτυξη. Χρησιμοποιούμε τα δεδομένα των Barro και Lee που προέρχονται από μια ομάδα 138 χωρών κατά την περίοδο 1960 με 1985. Στόχος μας είναι να εξετάσουμε τους εθνικούς ρυθμούς αύξησης του κατά κεφαλήν ΑΕΠ [7] (*Gross Domestic Product (GDP) per capita*), ως την εξαρτημένη μεταβλητή y, για τις περιόδους 1965-1975 και 1975-1985. Γενικά, ο ρυθμός ανάπτυξης του ΑΕΠ σε μια περίοδο από t_1 μέχρι t_2 ορίζεται συνήθως ως $log(GDP_{t_2}/GDP_{t_1})$.

Στην ανάλυση μας θα εξετάσουμε ένα μοντέλο με $p = 62$ συμμεταβλητές, το οποίο επιτρέπει συνολικά $n = 90$ πλήρεις παρατηρήσεις. Ο στόχος μας εδώ είναι να επιλέξουμε ένα υποσύνολο αυτών των συμμεταβλητών μεταξύ των οποίων υπάρχουν μεταβλητές μέτρησης της εκπαίδευσης, των πολιτικών επιστημών, τη δύναμη των θεσμών αγοράς, το άνοιγμα του εμπορίου, τα ποσοστά αποταμίευσης και άλλα πολλά. Έπειτα η θεωρία προβλέπει ότι για τις χώρες με παρόμοια τα άλλα χαρακτηριστικά το αποτέλεσμα του αρχικού επίπεδου του ΑΕΠ για το ρυθμό ανάπτυξης θα πρέπει να είναι αρνητικό. Έτσι η μορφή του μοντέλου μας θα έχει την ακόλουθη μορφή:

$$y_i = a_0 + a_1 log G_i + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i \qquad (2.17)$$

όπου το y_i είναι ο ρυθμός ανάπτυξης του ΑΕΠ σε μια συγκεκριμένη δεκαετία της χώρας *i*, G_i είναι το αρχικό επίπεδο του ΑΕΠ στη αρχή της συγκεκριμένης

[7] Το Ακαθάριστο Εγχώριο Προϊόν είναι το σύνολο των προϊόντων, υλικών και άυλων, που παράχθηκαν μέσα στην επικράτεια μιας χώρας σε διάστημα ενός έτους, εκφρασμένο σε χρηματικές μονάδες, ακόμα και αν μέρος αυτού παράχθηκε από παραγωγικές μονάδες που ανήκουν σε κατοίκους του εξωτερικού. Το κατά κεφαλήν ΑΕΠ είναι το ΑΕΠ διαιρούμενο με το μέσο πληθυσμό του έτους. Με λίγα λόγια, το κατά κεφαλήν ΑΠ προσαρμοσμένο σε ισοτιμία αγοραστικής δύναμης είναι ένας ασφαλής τρόπος για να μετρήσει κανείς το πόσο πλούσια είναι μια χώρα.

περιόδου που βασίζεται η μελέτη, και X_{ij} είναι ένας κατάλογος των χαρακτηριστικώς της χώρας *i* στην αρχή της συγκεκριμένης περιόδου. Μας ενδιαφέρει ο έλεγχος της υπόθεσης της σύγκλισης, δηλαδή ότι $\alpha_1 < 0$.

Εφαρμόσαμε επιλογή συμμεταβλητής χρησιμοποιώντας την LASSO, όπου χρησιμοποιήσαμε την *data-driven* επιλογή ενός επιπέδου ποινής λ με δύο τρόπους. Πρώτα χρησιμοποιήσαμε ένα άνω φράγμα για το σ να είναι το $\hat{\sigma}^0$ και μειώσαμε την ποινή για την εκτίμηση διαφορετικών μοντέλων με $\lambda, \frac{\lambda}{2}, \frac{\lambda}{3}$,και $\frac{\lambda}{4}$. Δεύτερο, εφαρμόσαμε επαναληπτική διαδικασία για τον καθορισμό του λ^{it} που καθορίζεται με βάση το $\hat{\sigma}^{it}$ το οποίο με τη σειρά του λαμβάνεται χρησιμοποιώντας την διαδικασία των επαναλήψεων της Post-LASSO που περιγράφεται στην παράγραφο 2.2.4.1 .

Η αρχική επιλογή της πρώτης προσέγγισης, λ, μας οδήγησε στο να μην επιλέξουμε συμμεταβλητές, κάτι που έρχεται σε συμφωνία με την *over-regularization* δεδομένου ότι το άνω φράγμα του σ είχε χρησιμοποιηθεί. Στη συνέχεια προχωρήσαμε στη σταδιακή μείωση του επίπεδου ποινής, λ, προκειμένου να καταστεί δυνατή η επιλογή κάποιων συμμεταβλητών. Παρουσιάζουμε τα αποτελέσματα της επιλογής του μοντέλου στον Πίνακα 2.2.

Με την πρώτη χαλάρωση της επιλογής του λ σε $\lambda/2$ τα δύο χαρακτηριστικά-συμμεταβλητές που επιλέγονται είναι η συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά (*black market exchange rate premium*)[8] , που αποτελεί χαρακτηριστικό του εμπορίου, και ένα μέτρο της πολιτικής αστάθειας[9]. Με την δεύτερη χαλάρωση της επιλογής του λ σε $\lambda/3$ επιλέγεται ένα επιπρόσθετο σύνολο των μεταβλητών που αντιστοιχεί στο λόγο της εικονικής κρατικής δαπάνης για άμυνα προς το εικονικό ΑΕΠ (*ratio of nominal government expenditure on defense to nominal GDP*) και στην αναλογία των εισαγωγών[10] προς το ΑΕΠ

[8] Η *black market premium* είναι η διαφορά μεταξύ της αξίας του νομίσματος στη μαύρη αγορά (παράνομο σύστημα) και της επίσημης συναλλαγματικής του ισοτιμίας (τιμή) σε σχέση με ένα άλλο νόμισμα.

[9] Είναι η πιθανότητα οργάνωσης διαδηλώσεων, εργαζόμενοι να πηγαίνουν σε διαδηλώσεις ή ακόμη και η δυνατότητα ύπαρξης ενός πραξικοπήματος. Μετριέται επίσης από την άποψη αν η κυβέρνηση μπορεί να καταρρεύσει ή όχι. Συνήθως η πολιτική αστάθεια συνδέεται με την έννοια μιας χρεοκοπημένης χώρας.

[10] Εισαγωγή ονομάζεται η διαδικασία μεταφοράς προϊόντων, εμπορευμάτων και ανθρώπων από μία ξένη χώρα σε μία άλλη όπου τα προϊόντα θα επεξεργαστούν, θα χρησιμοποιηθούν, θα πωληθούν ή θα επανεξαχθούν.

(*ratio of import to GDP*). Μέχρι τώρα καθώς μειώναμε το επίπεδο ποινής λ παρατηρούμε ότι αυξάνεται ο αριθμός των συμμεταβλητών. Παρόλα αυτά δεν αποτελεί κανόνα μιας και στην μείωση του σε $\lambda/4$ αφαιρείται η παράμετρος του λόγου των εισαγωγών προς το ΑΕΠ και μένουν οι υπόλοιπες 3 συμμεταβλητές.

Από την άλλη με το δεύτερο τρόπο, της επαναληπτικής προσέγγισης καταλήξαμε σε ένα μοντέλο με μια μόνο συμμεταβλητή που είναι η συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά.

Στη συνέχεια προχωρήσαμε στην εφαρμογή απλής γραμμικής παλινδρόμησης στα επιλεγόμενα μοντέλα και υπολογίστηκαν επίσης τα τυποποιημένα διαστήματα εμπιστοσύνης για αυτές τις εκτιμήσεις. Τα αποτελέσματα παρουσιάζονται στο Πίνακα 2.1. Διαπιστώνουμε ότι σε όλα τα μοντέλα με επιπλέον επιλεγόμενες συμμεταβλητές, οι συντελεστές γραμμικής παλινδρόμησης στο αρχικό επίπεδο του ΑΕΠ είναι πάντα αρνητικοί και τα τυποποιημένα διαστήματα εμπιστοσύνης δεν περιλαμβάνουν το μηδέν.

Το συμπέρασμα είναι ότι αυτά τα εμπειρικά ευρήματα υποστηρίζουν σθεναρά την υπόθεση της σύγκλισης (*conditional convergence*) που προέρχεται από το κλασσικό μοντέλο ανάπτυξης Solow-Swan-Ramsey. Η συγκεκριμένη υπόθεση αναφέρει ότι οι φτωχότερες χώρες θα πρέπει να αναπτύσσονται γρηγορότερα και ως εκ τούτου θα πρέπει να τείνουν να καλύψουν την διαφορά με τις πλουσιότερες χώρες. Μια τέτοια υπόθεση συνεπάγει ότι η επίδραση του αρχικού επιπέδου του ΑΕΠ για το ρυθμό ανάπτυξης θα πρέπει να είναι αρνητική. Όπως επισημάνουν Barro και Sala-i-Martin, αυτή η υπόθεση απορρίπτεται χρησιμοποιώντας την απλή διμεταβλητή παλινδρόμηση (*bivariate regression*) των ρυθμών ανάπτυξης στο αρχικού επίπεδου του ΑΕΠ.

Πίνακας 2.1: Ο παραπάνω πίνακας παρουσιάζει τον συντελεστή και ένα διάστημα εμπιστοσύνης 90% για κάθε μοντέλο που επιλέγεται από το αντίστοιχο επίπεδο ποινής.

Διαστήματα Εμπιστοσύνης μετά την επιλογή μοντέλων για την παλινδρόμηση Cross-Country Growth		
Παράμετρος ποινικοποίησης	Πραγματικό κατά κεφαλήν ΑΕΠ(λογαριθμικό)	
λ=2.7870	Συντελεστής	Διάστημα εμπιστοσύνης 90%
$\lambda^{it} = 2.3662$	-0.0112	[-0.0219,-0.0007]
$\lambda/2$	-0.0120	[-0.0225,-0.0015]
$\lambda/3$	-0.0153	[-0.0261,-0.045]
$\lambda/4$	-0.0221	[-0.0346,-0.0097]

Πίνακας 2.2: Τα επιλεγόμενα μοντέλα για τα διάφορα επίπεδα ποινής

Αποτελέσματα της επιλογής μοντέλου για την παλινδρόμηση Cross-Country Growth	
Παράμετρος ποινικοποίησης	**Επιλεγμένες συμμεταβλητές για το μοντέλο μας**
λ	-
λ^{it}	Συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά (log)
$\lambda/2$	Συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά (log) Μέτρο πολιτικής αστάθειας
$\lambda/3$	Συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά (log) Μέτρο πολιτικής αστάθειας Λόγος της εικονικής κρατικής δαπάνης για άμυνα προς το εικονικό ΑΕΠ Αναλογία των εισαγωγών προς το ΑΕΠ
$\lambda/4$	Συναλλαγματική ισοτιμία της διαφοράς στη μαύρη αγορά (log) Μέτρο πολιτικής αστάθειας Λόγος της εικονικής κρατικής δαπάνης για άμυνα προς το εικονικό ΑΕΠ

2.3 Μέσο μοντέλο παλινδρόμησης

Ένα μεγάλο μέρος της βιβλιογραφίας των αραιών μοντέλων υψηλών διαστάσεων επικεντρώνεται στο μέσο μοντέλο παλινδρόμησης (*mean regression model*). Σε αυτή την παράγραφο θα γίνει περιγραφή μεθόδων που αφορούν την παλινδρόμηση ποσοστημορίου και τα γενικευμένα γραμμικά μοντέλα στην περίπτωση των αραιών υψηλών διαστάσεων.

2.3.1 Παλινδρόμηση Ποσοστημορίου

Η παλινδρόμηση στο ποσοστημόριο (*quantile regression***)**, σύμφωνα με την εργασία των Koenker και Basett (1978), μπορεί να θεωρηθεί ως η φυσική εξέλιξη

της κλασσικής εκτίμησης των ελάχιστων τετραγώνων της πολλαπλής παλινδρόμησης, στην εκτίμηση ενός συνόλου μοντέλων με δεσμευμένες συναρτήσεις ποσοστημορίων.

2.3.1.1 Γενικά για ποσοστημόρια και Βελτιστοποίηση

Προτού αναφερθούμε στην παλινδρόμηση ποσοστημορίου σχετικά με τις μεθόδους ποινικοποίησης που αναφέραμε πιο πάνω, θα περιγράψουμε κάποια γενικά βασικά στοιχεία που αφορούν και συμβάλουν στην εφαρμογή της παλινδρόμησης στα ποσοστημόρια.

Ειδικότερα, έστω ότι X μια τυχαία μεταβλητή με συνάρτηση κατανομής:

$$F(x) = P(X \leq x) \tag{2.18}$$

όπου για κάθε $0 < \tau < 1$

$$Q_y(\tau) = F^{-1}(\tau) = inf(x: F(x) \geq \tau) \tag{2.19}$$

καλείται το τ-οστό ποσοστημόριο της X. Τα ποσοστημόρια μπορούν να προκύψουν από την λύση ενός προβλήματος βελτιστοποίησης.

Έστω ότι η απώλεια περιγράφεται από τη συνάρτηση:

$$p_\tau(u) = u\big(\tau - 1(u < 0)\big) \tag{2.20}$$

Για κάποιο $\tau \in (0,1)$, τότε ζητείται το u που ελαχιστοποιεί τη ζημιά. Ψάχνουμε λοιπόν να ελαχιστοποιήσουμε τη

$$E_{\rho_\tau}(X - \hat{x}) = (\tau - 1) \int_{-\infty}^{\hat{x}} (x - \hat{x}) dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x}) dF(x) \tag{2.21}$$

Παραγωγίζοντας ως προς $\hat{x}$, έχουμε ότι

$$(\tau - 1)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = 0 \;\Rightarrow\; F(\hat{x}) - \tau = 0 \;\Rightarrow\; F(\hat{x}) = \tau \qquad (2.22)$$

Αφού η F είναι μονότονη, οποιοδήποτε στοιχείο του $\{x : F(x) = \tau\}$ ελαχιστοποιεί την αναμενόμενη απώλεια. Όταν έχουμε μοναδική λύση τότε $\hat{x} = F^{-1}(\tau)$, διαφορετικά έχουμε ένα διάστημα τ-ποσοστημορίων από τα οποία το μικρότερο στοιχείο πρέπει να επιλεχθεί για να τηρήσουν την υπόθεση ότι η εμπειρική συνάρτηση των ποσοστημορίων είναι συνεχής από αριστερά.

Όταν η συνάρτηση κατανομής F αντικαθίσταται από την εμπειρική συνάρτηση κατανομής

$$F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \le x) \qquad (2.23)$$

επιλέγουμε το $\hat{x}$ για την ελαχιστοποίηση της αναμενόμενης ζημιάς:

$$\int \rho_\tau(x - \hat{x}) dF_n(x) = n^{-1}\sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}) \qquad (2.24)$$

Και έτσι θα παράγουμε το τ-οστό δειγματικό ποσοστημόριο.

Για να καταλάβουμε πως λειτουργεί η παραπάνω διαδικασία αρκεί να δώσουμε ένα απλό παράδειγμα. Έστω ότι έχουμε την διακριτή τυχαία μεταβλητή Y που παίρνει τις τιμές $1, \dots, 9$ με την ίδια πιθανότητα και θέλουμε να βρούμε την διάμεσο της Y. Υποθέτουμε επίσης, $\tau = 0.5$ και $u = 3$. Σύμφωνα λοιπόν με τον παραπάνω τύπο, θα έχουμε ότι η αναμενόμενη απώλεια υπολογίζεται από τον τύπο

$$\frac{\tau-1}{9}\sum_{y_i<u}(y_i-u)\ +\frac{\tau}{9}\sum_{y_i\geq u}(y_i-u) \qquad (2.25\ \alpha)$$

Πραγματοποιώντας τις πράξεις καταλήγουμε στο αποτέλεσμα (επειδή το $\tau/9$ είναι μια σταθερά, μπορεί να παραληφθεί από την παραπάνω αναμενόμενη συνάρτηση απώλειας)

$$\sum_{i=1}^{2}-(i-3)+\sum_{i=3}^{9}(i-3)=[(2+1)+(0+1+2+\cdots 6)]=24 \qquad (2.25\ \beta)$$

Αν αυξήσουμε τώρα το u κατά μια μονάδα, το αποτέλεσμα της παραπάνω εξίσωσης θα μειωθεί κατά 3 μονάδες. Στην περίπτωση όμως που αυξησουμε το u κατά δύο μονάδες, δηλαδή $u=5$, το αποτέλεσμα που θα πάρουμε είναι $\sum_{i=1}^{4}-(i-5)+\ \sum_{i=0}^{4}(i-5)=20$. Αφού υπολογίσουμε την εξίσωση για όλες τις διακριτές τιμές της τυχαίας μεταβλητής Y καταλήγουμε στο συμπέρασμα ότι το $u=5$ είναι η διαμεσος. Ο παρακάτω πίνακας παρουσιάζει για κάθε τιμή της u το αποτέλεσμα της εξίσωσης.

Πίνακας 2.3: Αποτελέσματα της εξίσωσης (2.25 β) για κάθε τιμή του u

u	1	2	3	4	5	6	7	8	9
Αποτέλεσμα	36	29	24	21	20	21	24	29	36

Με τον τρόπο αυτό εκφράσαμε το πρόβλημα της εύρεσης του δειγματικού ποσοστημορίου, που φαίνεται να είναι συνδεδεμένο με την έννοια της διάταξης των παρατηρήσεων του δείγματος ως τη λύση ενός απλού προβλήματος βελτιστοποίησης ή πιο συγκεκριμένα ελαχιστοποίησης. Στην πραγματικότητα έχουμε αντικαταστήσει την διαδικασία ταξινόμησης μέσω της διαδικασίας της βελτιστοποίησης.

Το πρόβλημα της εύρεσης του τ-οστού ποσοστημορίου, μπορεί να γραφεί ως εξής:

$$min_{\xi \in R} \sum_{i=1}^{n} \rho_\tau (y_i - \xi) \qquad (2.26)$$

Επομένως τα ποσοστημόρια εκφράζονται ως η λύση ενός προβλήματος βελτιστοποίησης και συνεπώς με αντίστοιχο τρόπο μπορούμε να προβούμε στην εκτίμηση μοντέλου συναρτήσεων με δεσμευμένα ποσοστημόρια.

2.3.1.2 Παλινδρόμηση Ποσοστημορίου – Μέθοδος LASSO

Αφού έγινε μια σύντομη αναφορά στην έννοια των ποσοστημορίων θα συνεχίσουμε σε πιο συγκεκριμένες περιπτώσεις, που αφορούν το αντικείμενο που μελετάμε.

Αρχικά, θεωρούμε ότι έχουμε τη μεταβλητή απόκρισης y_i καθώς και τις p-διαστάσεων επεξηγηματικές μεταβλητές x_i έτσι ώστε η u -οστή εξαρτημένη συνάρτηση ποσοστημορίου του y_i δεδομένων των x_i να δίνεται από την ακόλουθη σχέση:

$$F^{-1}_{y_i|x_i}(u|x) = x'\beta(u), \ \ \mu\varepsilon\, \beta(u) \ \in \ R^p \qquad (2.27)$$

όπου το $u \in (0,1)$ είναι ο δείκτης ποσοστημορίου που μας ενδιαφέρει .

Να σημειωθεί ότι η u-οστή εξαρτημένη συνάρτηση ποσοστημορίου $F^{-1}_{y_i|x_i}(u|x)$ είναι η αντίστροφη της εξαρτημένης συνάρτησης $F^{-1}_{y_i|x_i}(y|x)$ του y_i δεδομένου του $x_i = x$. Στη συνέχεια υποθέτουμε ότι το πραγματικό μοντέλο $\beta(u)$ έχει τον ακόλουθο αραιό φορέα:

$$T_u = support\big(\beta(u)\big) = \big\{ j \ \in \{1, \dots, p\} : \ \big|\beta_j(u)\big| > 0 \ \big\} \qquad (2.28)$$

Το συγκεκριμένο στήριγμα έχει $s_u \leq s \leq n/\log(n \vee p)$ μη μηδενικά στοιχεία.

Όπως είναι γνωστό, ο συντελεστής $\beta(u)$ συμβάλλει στην ελαχιστοποίηση της συνάρτησης κριτηρίου

$$Q_u(\beta) = E_n[\rho_u(y_i - x_i'\beta)] \quad (2.29)$$

όπου $\rho_u(t) = (u - 1\{t \leq 0\})t$ είναι η ασύμμετρη συνάρτηση απόλυτης απόκλισης.

Δεδομένου του τυχαίου δείγματος $(y_1, x_1),\ldots, (y_n, x_n)$ ο εκτιμητής παλινδρόμησης του ποσοστημορίου $\hat{\beta}(u)$ ορίζεται ως αυτός που θα ελαχιστοποιήσει την:

$$\hat{Q}_u(\beta) = E_n[\rho_u(y_i - x_i'\beta)] \quad (2.30)$$

Εξαιτίας του γεγονότος ότι η παλινδρόμηση του ποσοστημορίου που εξετάζουμε είναι μη συνεπής, απαιτείται να γίνει χρήση της ποινικοποίησης έτσι ώστε να εξαφανιστούν οι ανεξάρτητες μεταβλητές των οποίων οι συντελεστές είναι μηδενικοί. Ως απόρροια του προηγούμενου θα έχουμε ότι πλέον ο l_1-ποινικοποιημένος εκτιμητής παλινδρόμησης του ποσοστημορίου, $\hat{\beta}(u)$ θα αποτελεί λύση του παρακάτω προβλήματος βελτιστοποίησης:

$$min_{\beta \in R^\rho} \hat{Q}_u(\beta) + \frac{\lambda\sqrt{u(1-u)}}{n} \|\beta\|_1 \quad (2.31)$$

Σε αυτό το σημείο να σημειωθεί ότι η συνάρτηση κριτηρίου της σχέσης (2.31) αποτελεί το άθροισμα των συναρτήσεων κριτηρίων της σχέσης (2.30) ενώ η συνάρτηση ποινής δίνεται από την κλιμακωτή l_1-νόρμα του παραμετρικού διανύσματος.

Σχετικά με το επίπεδο ποινής, λ , ο προσδιορισμός του θα γίνει με τη βοήθεια της ακόλουθης τυχαίας μεταβλητής Λ :

$$\Lambda = n\, max_{1 \leq j \leq p} \left| E_n \left[\frac{x_{ij}(u - 1\{u_i \leq u\})}{\sqrt{u(1-u)}} \right] \right| \quad (2.32)$$

όπου οι $u_1, \dots, u_n$ είναι τυχαίες μεταβλητές ομοιόμορφα κατανεμημένες στο $(0,1)$ και ανεξάρτητα κατανεμημένες από τις ανεξάρτητες μεταβλητές $x_1, \dots x_n$. Από την άλλη η τυχαία μεταβλητή Λ έχει εξαρτημένη κατανομή στο σύνολο $X = [x_1, \dots x_n]'$.

Έτσι, με τη χρήση της τυχαίας μεταβλητής Λ οι Belloni και Chernozhukov (2011) όρισαν το επίπεδο ποινής ως:

$$\lambda = c\Lambda(1-\gamma|X) \qquad (2.33)$$

όπου $\Lambda(1-\gamma|X) \coloneqq (1-\gamma) -$ ποσοστημόριο του Λ εξαρτημένο στο X και $1-\gamma$ είναι το επίπεδο εμπιστοσύνης που απαιτείται να πλησιάζει όσο το δυνατό περισσότερο την μονάδα.

Τέλος, ο post-QR ποινικοποιημένος εκτιμητής (*post-l_1-QR*) εφαρμόζει την παλινδρόμηση του ποσοστημορίου στο μοντέλο $\hat{T}_u$ που επιλέγεται από την l_1-ποινικοποιημένη παλινδρόμηση του ποσοστημορίου.

Ειδικότερα, το μοντέλο και ο post-QR ποινικοποιημένος εκτιμητής καθορίζονται ως:

$$\hat{T}_u = support\left(\hat{\beta}(u)\right) = \{j \in \{1, \dots, p\} : \left|\hat{\beta}_j(u)\right| > 0 \} \qquad (2.34\,\alpha)$$

$$\tilde{\beta}(u) \in argmin\, \hat{Q}_u(\beta) : \beta_j = 0, j \in \hat{T}^c \qquad (2.34\,\beta)$$

όπου η τελευταία σχέση είναι η παλινδρόμηση ποσοστημορίου αφαιρώντας τις ανεξάρτητες μεταβλητές που δεν είχαν επιλεγεί στο πρώτο βήμα.

2.3.2 Γενικευμένα Γραμμικά Μοντέλα

Είναι πλέον εμφανές ότι οι l_1-κανονικοποιημένες μέθοδοι έχουν την δυνατότητα να εφαρμοσθούν και σε άλλες συναρτήσεις κριτηρίων πέραν των ελάχιστων

τετραγώνων και της παλινδρόμησης ποσοστημορίου. Συγκεκριμένα ο Sara van de Geer (2008) τις χρησιμοποίησε στα γενικευμένα γραμμικά μοντέλα.

Καταρχήν, έστω ότι έχουμε την μεταβλητή απόκρισης $y \in R$ καθώς και τις συμμεταβλητές (*covariates*) $x \in R^p$. Τότε η συνάρτηση κριτηρίου σε αυτή την περίπτωση ορίζεται ως εξής:

$$\hat{Q}(\beta) = \frac{1}{n}\sum_{i=1}^{n} h(y_i, x_i'\beta) \tag{2.35}$$

όπου η h είναι διαφορίσιμη με παράγωγο $\forall h$, κυρτή και 1-*Lipschitz* όσον αφορά το δεύτερο όρισμα, $|h(y,t) - h(y,t')| \leq |t - t'|$.

Σχετικά με την παράμετρο του πραγματικού μοντέλου θα έχουμε ότι:

$$\beta_0 \in argmin_{\beta \in R^\rho} E[h(y_i, x_i{}'\beta)] \Rightarrow \; E[x_i \nabla h(y_i, x_i'\beta)] = 0 \tag{2.36}$$

Ο l_1-ποινικοποιημένος εκτιμητής θα είναι η λύση του ακόλουθου προβλήματος:

$$min_{\beta \in R^\rho} \hat{Q}(\beta) + \frac{\lambda}{n}\|\beta\|_1 \tag{2.37}$$

Η επιλογή της παραμέτρου ποινής λ, σύμφωνα με τον Sara van de Geer (2008) βασίζεται στη χρήση των ανισοτήτων του Ledoux και Talagrand (1991) προκειμένου να δεσμεύεται ο βαθμός:

$$n\left\|\nabla \hat{Q}(\beta_0)\right\|_\infty = \|\sum_{i=1}^{n} x_i \nabla h(y_i, x_i'\beta_0)\|_\infty \lesssim_P \|\sum_{i=1}^{n} x_i \xi_i\|_\infty \tag{2.38}$$

όπου ξ_i είναι οι ανεξάρτητες τυχαίες μεταβλητές Rademacher για τις οποίες ισχύει ότι $P(\xi_i = 1) = P(\xi_i = -1) = 1/2$

Ακολούθως προτάθηκαν από τον Sara van de Geer (2008) περαιτέρω φράγματα στην δεξιά πλευρά της σχέσης (2.38). Για λόγους αποτελεσματικότητας,

προτείνεται η προσομοίωση $1-\gamma$ των ποσοστημορίων στην δεξιά πλευρά της σχέσης (2.38) που εξαρτάται από τις ανεξάρτητες μεταβλητές. Σε κάθε περίπτωση μπορεί να επιτευχθεί η επικράτηση του θορύβου $\lambda/n \geq c\left\|\nabla\hat{Q}(\beta_0)\right\|_\infty$ με μεγάλη πιθανότητα. Και επίσης από το γεγονός ότι η h είναι 1-*Lipschitz*, η επιλογή του επίπεδου ποινής είναι καθοριστική.

ΚΕΦΑΛΑΙΟ 3

ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΚΤΙΜΗΣΗΣ ΓΙΑ ΤΑ ΑΡΑΙΑ ΜΟΝΤΕΛΑ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

3.1 Ρυθμός Σύγκλισης για τις μεθόδους LASSO και Post-LASSO

Σε αυτή την ενότητα ένα σημείο που αποτελεί πρωτεύουσας σημασίας είναι τα αποτελέσματα της μελέτης σχετικά με την εκτίμηση της β_0 η οποία έχει μόνο $s < n$ μη μηδενικές συνιστώσες. Θα γίνει χρήση των αποτελεσμάτων του ρυθμού σύγκλισης της νόρμας πρόβλεψης (*prediction norm*), η οποία μετρά την ακρίβεια της πρόβλεψης $x_i'\beta_0$ στα σημεία σχεδιασμού $x_1, \dots, x_n$ και δίνεται από τον ακόλουθο τύπο:

$$\|\delta\|_{2,n} := \sqrt{E_n[(x_i'\delta)^2]} = \sqrt{\delta' E_n[x_i x_i']\delta} \qquad (3.1)$$

Η πρόβλεψη δίνεται από τον τύπο $\delta := \hat{\beta} - \beta_0$, γεγονός που σημαίνει ότι ορίζει τις αποκλίσεις των εκτιμητών $\hat{\beta}$ από τη τιμή της πραγματικής παραμέτρου β_0. Επομένως, στην περίπτωση μας, η ποσότητα $\|\delta\|_{2,n}^2$ υποδηλώνει τον μέσο όρο των τετραγωνικών σφαλμάτων $x_i'\hat{\beta} - x_i'\beta_0$ που προκύπτει με την χρήση της εκτίμησης $x_i'\hat{\beta}$ αντί της $x_i'\beta_0$.

Να συμπληρώσουμε επίσης ότι μπορεί να δεσμευτεί και ο εμπειρικός κίνδυνος των προβλεπόμενων τιμών f_i από την εκτίμηση $x_i'\hat{\beta}$ μέσω της τριγωνικής ανισότητας :

$$\sqrt{E_n[x_i'\hat{\beta} - f_i\,]^2} \leq \left\|\hat{\beta} - \beta_0\right\|_{2,n} + c_s \qquad (3.2)$$

Όπως φαίνεται και από τον ορισμό της , η νόρμας πρόβλεψης εξαρτάται άμεσα από τον πίνακα Gram $E_n[x_i x_i']$. Ωστόσο, τόσο στα παραμετρικά όσο και στα μη παραμετρικά μοντέλα υψηλών διαστάσεων, ο πίνακας αυτός δεν είναι πλήρους βαθμού (*full rank*) [11] και άρα δεν παρουσιάζει καλή συμπεριφορά. Αυτό που μας ενδιαφέρει όμως είναι η σωστή συμπεριφορά ορισμένων συντελεστών αυτού του πίνακα που ονομάζονται αραιές ιδιοτιμές (*sparse eigenvalues*).

Ακολούθως, θα οριστεί η ελάχιστη και μέγιστη αραιή ιδιοτιμή του εμπειρικού πίνακα Gram $E_n[x_i x_i']$, οι οποίες θα χρησιμεύσουν σε Θεωρήματα καθώς και στην ανάλυση σχετικά με τις εκτιμήσεις.

Ελάχιστη αραιή ιδιοτιμή :

$$\varphi_{min}(m)\left[E_n[x_i x_i']\right] \coloneqq min_{\|\delta\|_0 \le m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2} = min_{\|\delta\|_0 \le m, \delta \neq 0} \frac{\delta' E_n[x_i x_i']\delta}{\|\delta\|^2} \qquad (3.3)$$

Μέγιστη αραιή ιδιοτιμή:

$$\varphi_{max}(m)\left[E_n[x_i x_i']\right] \coloneqq max_{\|\delta\|_0 \le m, \delta \neq 0} \frac{\|\delta\|_{2,n}^2}{\|\delta\|^2} = max_{\|\delta\|_0 \le m, \delta \neq 0} \frac{\delta' E_n[x_i x_i']\delta}{\|\delta\|^2} \qquad (3.4)$$

Να σημειωθεί ότι το m αντιστοιχεί στο άνω φράγμα του αριθμού των μη μηδενικών συνιστωσών εκτός του φορέα T. Επίσης, υποθέτουμε ότι σχετικά με την ελάχιστη αραιή ιδιοτιμή ισχύει ότι $\varphi_{min}(m)\left[E_n[x_i x_i']\right] > 0$, γεγονός που σημαίνει ότι οι εμπειρικοί υποπίνακες Gram που σχηματίζονται από οποιεσδήποτε συνιστώσες m των x_i , εκτός βέβαια από τις συνιστώσες του T, είναι θετικά ορισμένοι.

[11] Ένας $\mu \times \nu$ πίνακας A ονομάζεται πλήρους βαθμού ακριβώς όταν $rankA = \min\{\mu, \nu\}$. Σε περίπτωση που δεν έχει πλήρη βαθμό τότε έχουμε το πρόβλημα της πολυσυγγραμμικότητας (είναι η κατάσταση η οποία δημιουργείται όταν υπάρχουν ισχυρές συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών στην πολλαπλή παλινδρόμηση).

ΣΥΝΘΗΚΗ SE

Υπάρχει $l_n \to \infty$ έτσι ώστε να ισχύει:

$$0 \le k \le \varphi_{min}(l_n s)\left[E_n[x_i x_i']\right] \le \varphi_{max}(l_n s)\left[E_n[x_i x_i']\right] \le k' \le \infty \qquad (3.5)$$

όπου k και k' είναι σταθερές οι οποίες είναι ανεξάρτητες από το μέγεθος του δείγματος n . Η συγκεκριμένη συνθήκη εφαρμόζεται πολύ καλά σε διάφορους σχεδιασμούς. Ειδικότερα, η συνθήκη *SE* ισχύει με πιθανότητα που συγκλίνει στη μονάδα καθώς το $n \to \infty$ εάν το x_i είναι η κανονικοποιημένη μορφή του $\tilde{x}_i$, δηλαδή $x_{ij} = \tilde{x}_{ij} / \sqrt{E_n\left[\tilde{x}_{ij}^2\right]}$, και:

- $\tilde{x}_i$, $i = 1, .., n$ είναι τυχαία γκαουσιανά διανύσματα με μηδενική μέση τιμή που αποτελούν πληθυσμό στο Gram πίνακα $E_n\left[\tilde{x}_{ij}\tilde{x}_{ij}'\right]$ με άσσους διαγώνια και οι ελάχιστες και μέγιστες $slogn$ −αραιές ιδιοτιμές είναι φραγμένες μακριά από το μηδέν και από πάνω και $slogn = o\left(\frac{n}{logp}\right)$.
- $\tilde{x}_i$, $i = 1, .., n$ είναι φραγμένα τυχαία διανύσματα με μηδενική μέση τιμή και $\|\tilde{x}_i\|_\infty \le K_n$ που αποτελούν πληθυσμό στο Gram πίνακα $E_n\left[\tilde{x}_{ij}\tilde{x}_{ij}'\right]$ με άσσους διαγώνια και οι ελάχιστες και μέγιστες $slogn$ −αραιές ιδιοτιμές είναι φραγμένες μακριά από το μηδέν και από πάνω και $K_n^2 slog^5(p \vee n) = o(n)$.

Παρατηρήσεις 3.1

Γενικά, στην οικονομετρία υποθέτουμε ότι ο πληθυσμός του πίνακα Gram $E[x_i x_i']$ έχει ιδιοτιμές οι οποίες είναι άνω και κάτω φραγμένες. Οι πιο πάνω προϋποθέσεις απαιτούν ότι μόνο οι αραιές $slogn$-ιδιοτιμές του συγκεκριμένου πίνακα είναι άνω και κάτω φραγμένες. Αυτό αποτελεί πρωτεύουσας σημασίας μιας και με τον τρόπο αυτό επιτρέπεται στις συναρτήσεις x_i να σχηματίζονται ως ένας συνδυασμός στοιχείων από διαφορετικές βάσεις· π.χ ένας συνδυασμός των B-*splines* με πολυώνυμα.

Θεώρημα 3.1

Ρυθμός σύγκλισης του εφικτού εκτιμητή LASSO $\widehat{\boldsymbol{\beta}}$

Υποθέτουμε ότι ισχύουν οι συνθήκες *ASM* και *SE* που περιεγράφηκαν προηγουμένως. Τότε για μεγάλο αριθμό δείγματος n ισχύουν τα ακόλουθα φράγματα με πιθανότητα τουλάχιστον $1-\gamma$:

$$C'\|\hat{\beta}-\beta_0\| \leq \|\hat{\beta}-\beta_0\|_{2,n} \leq C\sigma\sqrt{\frac{slog(2p/\gamma)}{n}} \qquad (3.6)$$

όπου $C>0$ και $C'>0$ είναι σταθερές, $C' \gtrsim \sqrt{\kappa'}$ και $C \lesssim \frac{1}{\sqrt{\kappa'}}$

και $log(p/\gamma) \lesssim log(p \vee n)$.

■

Παρατηρήσεις 3.2

Το Θεώρημα 3.1 αναφέρεται στο ρυθμό σύγκλισης του εφικτού εκτιμητή LASSO σε γκαουσιανο μοντέλο. Ο ρυθμός εκτίμησης του β_0 είναι το ριζικό του κλάσματος του αριθμού των παραμέτρων s στο πραγματικό μοντέλο, T, προς το μέγεθος του δείγματος , δηλαδή $\sqrt{s/n}$, πολλαπλασιασμένο με τον λογαριθμικό παράγοντα $\sqrt{\log(p \vee n)}$. Αυτός ο λογαριθμικός παράγοντας μπορεί να θεωρηθεί ως η τιμή του αγνώστου στο πραγματικό μοντέλο. Να σημειωθεί ότι ο ρυθμός εκτίμησης για την συνάρτηση παλινδρόμησης f πάνω στα σημεία σχεδιασμού προκύπτει από τον συνδυασμό της τριγωνικής ανισότητας και της συνθήκης *ASM*:

$$\sqrt{E_n\left[\left(f(z_i)-x_i'\hat{\beta}\right)^2\right]} \leq \|\hat{\beta}-\beta_0\|_{2,n} + c_s \lesssim_P \sigma\sqrt{\frac{s\,log(p \vee n)}{n}} \qquad (3.7)$$

Τα αποτελέσματα του Θεωρήματος 3.1 αποτελούν μια επέκταση των αποτελεσμάτων του θεμελιώδες έργου των Buickel, Ritov και Tsybakov (2009) και των

Meinshaunsen και Yu (2009) στον εφικτό LASSO και των Candès και Tao (2007) στον Dantzig selector.

Τα όρια του Θεωρήματος 3.1 σχετικά με την σύγκλιση του εφικτού εκτιμητή LASSO, επιτρέπουν τις κατασκευές συνόλων εμπιστοσύνης για το β_0, σύμφωνα με τους Chernozhukov (2009) και Gautier και Tsybakov (2011). Τέτοια σύνολα εμπιστοσύνης βασίζονται στα αποτελεσματικά φράγματα των C. Ο υπολογισμός των φραγμάτων για τα C απαιτεί οπωσδήποτε των υπολογισμό των συνδυασμένων ποσοτήτων βασιζόμενες στο άγνωστο μοντέλο T γεγονός το οποίο κάνει δύσκολο τον υπολογισμό της προσέγγισης.

Όπως γνωρίζουμε, οι l_1-κανονικοποιημένοι εκτιμητές έχουν μια εξ' ορισμού μεροληψία και για αυτό προτάθηκαν άλλωστε οι εκτιμητές Post-LASSO για να καταργήσουν αυτή την μεροληψία. Αποδεικνύεται ότι μπορούμε να δεσμεύσουμε την απόδοση του εκτιμητή Post-LASSO ως μια συνάρτηση του ρυθμού σύγκλισης LASSO και της δυνατότητας επιλογής μοντέλου LASSO. Για τους συνήθεις σχεδιασμούς, αυτή η δέσμευση δείχνει ότι η Post-LASSO μπορεί να εκτελείται τόσο καλά όσο και η LASSO και επιπλέον σε ορισμένες περιπτώσεις να πετυχαίνει και καλύτερα αποτελέσματα. Εκ κατασκευής είναι γνωστό ότι η εκτίμηση με Post-LASSO έχει μικρότερη μεροληψία σε σχέση με τη LASSO.

Θεώρημα 3.2

Ρυθμός σύγκλισης για τον εφικτό εκτιμητή Post-LASSO

Υποθέτοντας ότι ισχύουν οι συνθήκες του Θεωρήματος 3.1 και έστω ότι $\varepsilon > 0$. Τότε υπάρχουν σταθερές C' και C_ε έτσι ώστε με πιθανότητα $1 - \gamma$ να ισχύει

$$\hat{s} = \left|\hat{T}\right| \leq C's \tag{3.8}$$

και με πιθανότητα $1 - \gamma - \varepsilon$ να ισχύει

$$\sqrt{\kappa'}\left\|\tilde{\beta} - \beta_0\right\| \leq \left\|\tilde{\beta} - \beta_0\right\|_{2,n} \leq C_\varepsilon \sigma \sqrt{\frac{s\ log(p \vee n)}{n}} \tag{3.9}$$

Εάν επιπλέον $\left|\left\|\hat{\beta}\right\|_0 - s\right| = o(s)$ και $T \subseteq \hat{T}$ με πιθανότητα που πλησιάζει το ένα, τότε

$$\left\|\tilde{\beta} - \beta_0\right\|_{2,n} \lesssim_P \sigma\left[\sqrt{\frac{o(s)\ log(p \vee n)}{n}} + \sqrt{s/n}\right] \qquad (3.10)$$

Εάν $\hat{T} = T$ με πιθανότητα που πλησιάζει το ένα, τότε ο Post-LASSO επιτυγχάνει την απόδοση oracle

$$\left\|\tilde{\beta} - \beta_0\right\|_{2,n} \lesssim_P \sigma\sqrt{s/n} \qquad (3.11)$$

■

Παρατηρήσεις 3.3

Το Θεώρημα 3.2 ισχύει για τον εκτιμητή Post-LASSO $\tilde{\beta}$ που υπολογίζεται χρησιμοποιώντας το μοντέλο $\hat{T} = support(\hat{\beta})$ επιλεγόμενο από τον εφικτό εκτιμητή LASSO $\hat{\beta}$ που ορίστηκε στο Θεώρημα 3.1.

Μέσα από το Θεώρημα 3.2 φαίνεται ότι ο εφικτός εκτιμητής Post-LASSO επιτυγχάνει τον ίδιο oracle ρυθμό με τον εφικτό LASSO. Αυτό συμβαίνει παρά το γεγονός ότι ο εφικτός (*feasible*) εκτιμητής LASSO μπορεί να αποτύχει να επιλέξει σωστά το μοντέλο oracle T, δηλαδή να έχουμε ότι $T \nsubseteq \hat{T}$. Η παραπάνω διαπίστωση ισχύει αφού είναι απίθανο να είναι σημαντικές οι συνιστώσες του μοντέλου T που χάνει ο εφικτός εκτιμητής LASSO.

ΚΕΦΑΛΑΙΟ 4

ΒΟΗΘΗΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ ΥΨΗΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

4.1 Εισαγωγή

Γενικά γνωρίζουμε ότι το γραμμικό μοντέλο παλινδρόμησης χρησιμοποιείται στην εκτίμηση σχέσεων των μεταβλητών, δηλαδή κατά πόσο ένα σύνολο ανεξάρτητων μεταβλητών επηρεάζει την εξαρτημένη μεταβλητή ενός υποδείγματος. Μια από τις πιο κρίσιμες υποθέσεις για την εκτίμηση αφορά την ανεξαρτησία της επεξηγηματικής μεταβλητής με το σφάλμα του υποδείγματος, δηλαδή με τον διαταρακτικό όρο. Σε αυτή την περίπτωση οι επεξηγηματικές μεταβλητές χαρακτηρίζονται ως εξωγενείς, αφού θεωρούμε ότι δημιουργούνται εκτός δείγματος. Όταν όμως η εξωγένεια δεν ικανοποιείται τότε οι εκτιμητές θα χαρακτηρίζονται από μεροληψία και δεν θα είναι συνεπείς με αποτέλεσμα να οδηγηθούμε σε λάθος συμπεράσματα. Όταν η ασθενέστερη μορφή εξωγένειας δεν ικανοποιείται, δηλαδή όταν έστω και μια ανεξάρτητη μεταβλητή συσχετίζεται με το διαταρακτικό όρο, τότε έχουμε το πρόβλημα της ενδογένειας. Παραδείγματος χάριν παρουσιάζεται ενδογένεια λόγω σφάλματος μέτρησης των μεταβλητών ή εξαιτίας της ύπαρξης ταυτόχρονων εξισώσεων ή λόγω παράλειψης, σχετικών με το υπόδειγμα, ανεξάρτητων μεταβλητών κτλ. Το φαινόμενο της ενδογένειας είναι εξαιρετικά συνηθισμένο στα οικονομετρικά υποδείγματα.

Για την αντιμετώπιση του προβλήματος της ενδογένειας των ανεξάρτητων μεταβλητών χρησιμοποιούμε τη μέθοδο των βοηθητικών ή τεχνητών μεταβλητών (*Instrumental Variables method, IV*). Ο εκτιμητής των βοηθητικών μεταβλητών ικανοποιεί υπό προϋποθέσεις την ελάχιστη απαίτηση της συνέπειας και βασίζεται σε ένα σύνολο μεταβλητών εκτός των ενδογενών επεξηγηματικών μεταβλητών ή της μεταβλητής απόκρισης. Η φιλοσοφία της μεθόδου βασίζεται στην εύρεση κατάλληλων μεταβλητών που αφενός συσχετίζονται με τις «προβληματικές» ανεξάρτητες μεταβλητές και αφετέρου οριακά (δηλαδή όσο το μέγεθος του δείγματος τείνει στο άπειρο) δεν συσχετίζονται με το διαταρακτικό όρο του

υποδείγματος. Στη θεωρεία της στατιστικής ένας εκ των γνωστότερων εκτιμητών IV είναι αυτός των ελάχιστων τετραγώνων σε δύο στάδια (*TSLS*).

4.2 Μέθοδοι Βοηθητικών Μεταβλητών

Αφού έγινε μια σύντομη ανάλυση των βοηθητικών μεταβλητών θα τις συνδυάσουμε με τις μεθόδους που μελετάμε, LASSO και Post LASSO. Θα εξετάσουμε το γραμμικό μοντέλο βοηθητικών μεταβλητών (*IV*) με πολλές βοηθητικές μεταβλητές.

Καταρχήν, έστω ότι έχουμε το σύστημα εξισώσεων γκαουσιανού (*Gaussian*) μοντέλου:

$$y_{1i} = y_{2i}a_1 + w_i'a_2 + \zeta_i \tag{4.1}$$

$$y_{2i} = f(z_i) + \upsilon_i \tag{4.2}$$

$$\begin{pmatrix}\zeta_i \\ \upsilon_i\end{pmatrix} \mid z_i \sim N\left(0, \begin{pmatrix}\sigma_\zeta^2 & \sigma_{\zeta\upsilon} \\ \sigma_{\zeta\upsilon} & \sigma_\upsilon^2\end{pmatrix}\right) \tag{4.3}$$

όπου y_{1i} είναι η μεταβλητή απόκρισης, y_{2i} είναι ενδογενής μεταβλητή· δηλαδή η τιμής της θα προσδιορίζεται από το σύστημα , w_i είναι ένα k_w-διάνυσμα των μεταβλητών ελέγχου (*control variables*) [12], $z_i = (u_i', w_i')'$ είναι ένα διάνυσμα των βοηθητικών μεταβλητών, και (ζ_i, υ_i) είναι οι διαταραχές οι οποίες είναι ανεξάρτητες του z_i. Η συνάρτηση $f(z_i) = E[y_{2i}|z_i]$, δηλαδή η βέλτιστη μεταβλητή ελέγχου (*the optimal instrument*) , είναι μια άγνωστη και ενδεχομένως πολύ-

[12] Είναι μια μεταβλητή που διατηρείται σταθερή προκειμένου να εκτιμηθεί ή να διασαφηνιστεί η σχέση μεταξύ δύο άλλων μεταβλητών. Η μεταβλητή ελέγχου δεν πρέπει να συγχέεται σε καμία περίπτωση με την ελεγχόμενη μεταβλητή (*controlled variable*) που αποτελεί ένα εναλλακτικό ορισμό για την ανεξάρτητη μεταβλητή.

πλοκη συνάρτηση των στοιχειωδών βοηθητικών μεταβλητών z_i. Η βασική παράμετρος που μας ενδιαφέρει είναι ο συντελεστής του y_{2i}, το οποίου η πραγματική τιμή είναι το a_1. Επίσης, να σημειωθεί ότι τα $\{z_i\}$ αντιμετωπίζονται ως σταθερές.

Με τη βοήθεια αυτών που περιεγράφηκαν παραπάνω, κατασκευάζουμε ένα διάνυσμα υψηλών διαστάσεων των τεχνικών βοηθητικών μεταβλητών, το $x_i = P(z_i)$, με διάσταση p ενδεχομένως πολύ μεγαλύτερη από το μέγεθος του δείγματος n και με τη βοήθεια των συνθηκών που θα αναλυθούν παρακάτω. Ακολούθως, θα γίνει εκτίμηση της βέλτιστης μεταβλητής ελέγχου $f(z_i)$ με την βοήθεια του εκτιμητή

$$\hat{f}(z_i) = x_i'\hat{\beta} \tag{4.4}$$

όπου το $\hat{\beta}$ είναι ο εφικτός εκτιμητής LASSO ή Post-LASSO όπως ορίστηκε σε προηγούμενη ενότητα.

Οι αραιές μέθοδοι εκμεταλλεύονται την προσεγγιστική αραιότητα και εξασφαλίζουν ότι πολλά στοιχεία του $\hat{\beta}$ θα μηδενίζονται όταν το p είναι μεγάλο. Δηλαδή, οι αραιές μέθοδοι θα επιλέγουν ένα μικρό υποσύνολο των διαθέσιμων τεχνιτών βοηθητικών μεταβλητών.

Έστω ότι ορίζουμε το μέγεθος $A_i = (f(z_i), w_i')'$ να είναι το διάνυσμα της ιδανικής βοηθητικής μεταβλητής το οποίο να εκτιμάται από το

$$\widehat{A_\iota} = (\hat{f}(z_i), w_i')' \tag{4.5}$$

Επίσης, αφού ορίσουμε $d_i = (y_{2i}, w_i')'$, σχηματίζουμε τον εφικτό εκτιμητή IV χρησιμοποιώντας το διάνυσμα της εκτιμώμενης βοηθητικής μεταβλητής

$$\hat{\alpha}^* = \left(E_n\left[\hat{A}_i d_i'\right]\right)^{-1}\left(E_n\left[\hat{A}_i y_{1i}\right]\right) \tag{4.6}$$

Συνθήκη ASIV

Στο γραμμικό μοντέλο IV (4.1)-(4.3) με τις τεχνικές βοηθητικές μεταβλητές $x_i = P(z_i)$ ισχύουν οι ακόλουθες παραδοχές:

(α) οι τιμές των παραμέτρων σ_υ, σ_ζ και οι ιδιοτιμές του $Q_n = E_n[A_i A_i']$ είναι φραγμένες μακριά από το μηδέν και άνω ομοιόμορφα στο n.

(β) Η συνθήκη ASM ισχύει για το (4.2) , δηλαδή για κάθε $i = 1, \dots, n$ υπάρχουν $\beta_0 \in R^p$, έτσι ώστε $f(z_i) = x_i'\beta_0 + r_i$, $\|\beta_0\| \leq s$, $\{E_n[r_i^2]\}^{1/2} \leq K\sigma_\upsilon\sqrt{s/n}$, όπου η σταθερά K δεν εξαρτάται από το n.

(γ) Η συνθήκη SE ισχύει για $E_n[x_i x_i']$.

(δ) $s^2 log^2(p \vee n) = o(n)$.

■

Θεώρημα 4.1

Ασυμπτωτική κανονικότητα για εκτιμητές IV βασισμένους στη LASSO και Post-LASSO

Υποθέτουμε ότι ισχύει η συνθήκη ASIV. Ο εκτιμητής IV που δίνεται από την σχέση (4.6) είναι $\sqrt{n}$-συνεπής και είναι ασυμπτωτικά αποτελεσματικός , δηλαδή καθώς το n αυξάνεται :

$$\left(\sigma_\zeta^2 Q_n^{-1}\right)^{-\frac{1}{2}}\sqrt{n}\,(\hat{a}^* - a) = N(0,1) + o_p(1) \qquad (4.7)$$

και το αποτέλεσμα ισχύει για Q_n που δίνεται από $\hat{Q}_n = E_n[\hat{A}_i \hat{A}_i']$ και για σ_ζ^2 μέσα από $\hat{\sigma}_\zeta^2 = E_n[(y_{1i} - \hat{A}_i'\hat{a}^*)^2]$.

■

Παρατηρήσεις 4.1

Το Θεώρημα 4.1 δείχνει ότι ο εκτιμητής IV βασισμένος στην εκτίμηση του πρώτου βήματος με LASSO ή Post-LASSO είναι ασυμπτωτικά τόσο αποδοτικός όσο και ο μη εφικτός βέλτιστος εκτιμητής IV ο οποίος χρησιμοποιεί το $A_i = (f(z_i), w_i')'$ και επιτυγχάνει με αυτό τον τρόπο την ημιπαραμετρική απόδοση του φράγματος του Chamberlain (1987). Με την σειρά τους οι Belloni, Chernozhukov και Hansen (2010) δείχνουν ότι το αποτέλεσμα συνεχίζει να ισχύει όταν οι υπόλοιπες αραιές μέθοδοι χρησιμοποιούνται για να εκτιμήσουν τις βέλτιστες βοηθητικές μεταβλητές. Οι επαρκείς συνθήκες για να δείξουν ότι ο IV εκτιμητής, που αποκτήθηκε χρησιμοποιώντας αραιές μεθόδους για την εκτίμηση των βέλτιστων βοηθητικών μεταβλητών, είναι ασυμπτωτικά αποτελεσματικός περιλαμβάνουν τεχνικές συνθήκες και την συνθήκη $s^2 log^2(p \vee n) = o(n)$. Αυτός ο όρος απαιτεί οι βέλτιστες βοηθητικές μεταβλητές να είναι αρκετά ομαλές έτσι ώστε ένας σχετικά μικρός αριθμός των όρων της σειράς να μπορεί να χρησιμοποιηθεί για να τις προσεγγίσει καλά. Αυτή η ομαλότητα εξασφαλίζει ότι η επίδραση της εκτίμησης των εργαλείων στον εκτιμητή IV είναι ασυμπτωτικά αμελητέα. Ο όρος $s^2 log^2(p \vee n) = o(n)$ είναι σημαντικός και δεν μπορεί να είναι σημαντικά εξασθενημένος για το πλήρες δείγμα του εκτιμητή IV που περιεγράφηκε παραπάνω, Ωστόσο μπορούμε να αντικαταστήσουμε τον όρο αυτό με ένα ασθενέστερο όρο $slog(p \vee n) = o(n)$ όπως ορίζεται στους Belloni, Chernozhukov, και Hansen (2010). Επίσης οι Belloni, Chen, Chernozhukov, and Hansen (2010) απέδειξαν ότι το αποτέλεσμα του Θεωρήματος με μερικές κατάλληλες τροποποιήσεις, συνεχίζει να ισχύει και υπό την ετεροσκεδαστικότητα αν και ο εκτιμητής δεν επιτυγχάνει απαραιτήτως την ημιπαραμετρική απόδοση του φράγματος.

4.3 Ασθενής προσδιορισμός με πολλές βοηθητικές μεταβλητές

Θεωρούμε ότι έχουμε το ακόλουθο σύστημα εξισώσεων:

$$y_{1i} = y_{2i}a_1 + w_i'a_2 + \zeta_i, \quad \zeta_i \mid z_i \sim \mathrm{N}\,(0, \sigma_\zeta^2) \tag{4.8}$$

όπου y_{1i} είναι η μεταβλητή απόκρισης, y_{2i} είναι η ενδογενής μεταβλητή , w_i είναι ένα k_w-διάνυσμα των μεταβλητών ελέγχου, $z_i = (u_i', w_i')'$ είναι ένα διάνυσμα των βοηθητικών μεταβλητών (*IV*), και ζ_i είναι η διαταραχή η οποία είναι ανεξάρτητη του z_i. Τα $\{z_i\}$ αντιμετωπίζονται ως σταθερές.

Είναι προτιμότερο να χρησιμοποιήσουμε ένα διάνυσμα $x_i = P(z_i)$ των τεχνικών βοηθητικών μεταβλητών για το συμπέρασμα που είναι ισχυρό στον ασθενή προσδιορισμό. Προτείνουμε μια μέθοδο για τη συμπερασματολογία, βασισμένη στους ανατρέφοντες ελέγχους κατά σημείο (*inverting pointwise tests*), που εκτελείται χρησιμοποιώντας ένα στατιστικό sup-score το οποίο ορίζεται πιο κάτω. Η διαδικασία αυτή είναι παρόμοιου πνεύματος με αυτή των Anderson και Rubin (1949) και Staiger και Stock (1997) με την διαφορά ότι χρησιμοποιεί πολύ διαφορετικές στατιστικές οι οποίες είναι καλά προσαρμοσμένες στις περιπτώσεις με πολλές βοηθητικές μεταβλητές.

Προκειμένου να οριστεί το στατιστικό sup-score θα δούμε πρώτα την επίδραση των w_i ελέγχων στις βασικές μεταβλητές. Για ένα n-διάνυσμα $\{u_i, i = 1, \dots, n\}$, ορίζουμε $\bar{u}_i = u_i - w_i' E_n[w_i w_i']^{-1} E_n[w_i u_i]$. Επίσης, ορίζουμε $\tilde{x}_i = (\tilde{x}_{i1}, \dots, \tilde{x}_{ip})'$. Στον σχηματισμό που αναφερόμαστε , παραλείπουμε τα στοιχεία των w_i από $\tilde{x}_{ij}$ δεδομένου ότι απαλείφονται μέσω partialling out [13]. Έπειτα, κανονικοποιούμε χωρίς απώλεια της γενικότητας

$$E_n\left[\tilde{x}_{ij}^2\right] = 1 \quad , j = 1, \dots, p \tag{4.9}$$

Το στατιστικό sup-score για τη δοκιμή της υπόθεσης $\alpha_1 = \alpha$ παίρνει τη μορφή:

$$\Lambda_\alpha = max_{1 \le j \le p} \frac{\left| nE_n[(\tilde{y}_{1i} - \tilde{y}_{2i} a)\tilde{x}_{ij}] \right|}{\sqrt{E_n[(\tilde{y}_{1i} - \tilde{y}_{2i} a)^2 \tilde{x}_{ij}^2}} \tag{4.10}$$

[13] Είναι η διαδικασία κατά την οποία αφαιρούμε μια μεταβλητή δίνοντας της μια σταθερή τιμή, προκειμένου να προσδιοριστεί οποιοσδήποτε συσχετισμός μεταξύ άλλων μεταβλητών.

Εάν η υπόθεση $\alpha_1 = \alpha$ είναι πραγματική , τότε η κρίσιμη τιμή για την επίτευξη του επιπέδου γ είναι:

$$\Lambda(1-\gamma|W,X) = 1-\gamma - ποσοστημόριο\ του\ max_{1\le j\le p} \frac{\left|nE_n\left[\tilde{g}_i\tilde{x}_{ij}\right]\right|}{\sqrt{E_n\left[\tilde{g}_i^2\tilde{x}_{ij}^2\right]}} \mid W,X \qquad (4.11)$$

όπου $W = [w_1, \dots, w_n]'$, $X = [x_1, \dots, x_n]'$, και $g_1, \dots, g_n$ είναι κατανεμημένες μεταβλητές στην $N(0,1)$ ανεξάρτητες των W και X , όπου τα $\tilde{g}_i$ είναι τα αριστερά υπόλοιπα μετά την προβολή των $\{g_i\}$ στο $\{w_i\}$ όπως ορίστηκε πιο πάνω.

Μπορούμε να προσεγγίσουμε την κρίσιμη τιμή $\Lambda(1-\gamma|W,X)$ με σχετική προσομοίωση του X στο W. Επίσης , είναι δυνατό να χρησιμοποιηθεί ένα απλό ασυμπτωτικό φράγμα σε αυτή την κρίσιμη τιμή , δηλαδή

$$\Lambda(1-\gamma) \coloneqq c\sqrt{n}\Phi^{-1}(1-\gamma/2p) \le c\sqrt{2nlog(2p/\gamma)} \text{ για } c>1 \qquad (4.12)$$

Το πεπερασμένο δείγμα $(1-\gamma)$-περιοχή εμπιστοσύνης για το α_1 δίνεται από τη σχέση

$$C \coloneqq \{a \in R: \Lambda_\alpha \le \Lambda(1-\gamma|W,X)\} \qquad (4.13)$$

Καθώς το μεγάλο δείγμα $(1-\gamma)$-περιοχή εμπιστοσύνης δίνεται από τη σχέση

$$C \coloneqq \{a \in R: \Lambda_\alpha \le \Lambda(1-\gamma)\} \qquad (4.14)$$

Συνθήκη HDIV

Υποθέτουμε ότι ισχύει το γραμμικό μοντέλο IV (4.8). Θεωρούμε ένα p-διάνυσμα των βοηθητικών μεταβλητών $x_i = P(z_i), i = 1, \dots, n$, έτσι ώστε $(logp)/n \to 0$. Υποθέτουμε ότι οι ακόλουθες υποθέσεις ισχύουν ομοιόμορφα στο n:

(α) η παραμετρική τιμή σ_ζ είναι φραγμένη μακριά από το μηδέν και από πάνω.

(β) η διάσταση του w_i είναι φραγμένη και οι ιδιοτιμές του Gram πίνακα $E_n[w_i w_i']$ είναι φραγμένες μακριά από το μηδέν.

(γ) $\|w_i\| \leq K$ και $|\tilde{x}_{ij}| \leq K$ για όλα τα $1 \leq \iota \leq n$ και για όλα τα $1 \leq j \leq p$, όπου το K είναι μια σταθερά , ανεξάρτητη του δείγματος n.

Θεώρημα 4.2

Ένγκυρη Συμπερασματολογία βασισμένη στο στατιστικό Sup-score

(α) Υποθέτουμε ότι το γραμμικό μοντέλο IV (4.8) ισχύει.

Τότε $P(\alpha_1 \in C) = 1 - \gamma$.

(β) Επίσης, υποθέτουμε ότι η συνθήκη HDIV ισχύει.

Επομένως προκύπτει $P(\alpha_1 \in C') \geq 1 - \gamma - o(1)$.

(γ) Επιπλέον, εάν το α είναι τέτοιο ώστε

$$max_{1 \leq j \leq p} \frac{|\alpha - \alpha_1|\sqrt{n}\,|E_n[\tilde{y}_{2i}\tilde{x}_{ij}]|/\sqrt{logp}}{\sigma_\zeta + |\alpha - \alpha_1|\sqrt{E_n[\tilde{y}_{2i}^2 \tilde{x}_{ij}^2]}} \to \infty \qquad (4.15)$$

τότε $P(a \in C) = o(1)$ και $P(a \in C') = o(1)$

■

Παρατηρήσεις 4.2

Από το Θεώρημα 4.2 προκύπτει ότι οι περιοχές εμπιστοσύνης C και C' που κατασκευάστηκαν πιο πάνω έχουν πεπερασμένο δείγμα και μεγάλη ισχύ δείγματος αντίστοιχα. Επιπλέον, η πιθανότητα να συμπεριλαμβάνεται ένα εσφαλμένο σημείο α είτε στο C είτε στο C' τείνει στο μηδέν εφόσον το α είναι αρκετά απόμακρο από το α_1 και οι βοηθητικές μεταβλητές δεν είναι πολύ ασθενείς. Ιδίως, εάν υπάρχει μια ισχυρή βοηθητική μεταβλητή, οι περιοχές εμπιστοσύνης θα αποκλείσουν τελικά τα σημεία α που είναι περαιτέρω από $\sqrt{(logp)/n}$ μακριά

από το α_1. Επιπλέον, εάν υπάρχουν βοηθητικές μεταβλητές των οποίων ο συσχετισμός με την ενδογενή μεταβλητή είναι μεγαλύτερος από $\sqrt{(logp)/n}$, τότε οι περιοχές εμπιστοσύνης θα είναι ασυμπτωτικά φραγμένες. Να τονισθεί ότι ένα καλό χαρακτηριστικό της συγκεκριμένης κατασκευής είναι ότι παρέχει με αποδείξεις τις έγκυρες περιοχές εμπιστοσύνης και δεν απαιτεί τον υπολογισμό κάποιων συνδυαστικών ποσοτήτων. Τελικά, σημειώνουμε ότι δεν είναι δύσκολο να γενικευτούν τα αποτελέσματα για να καταστεί δυνατή η αύξηση του αριθμού των ελέγχων w_i κάτω από κατάλληλες τεχνικές προϋποθέσεις που περιορίζουν τον αριθμό των ελέγχων και την περιβάλλουσα (*envelope*)[14] τους σε σχέση με το μέγεθος του δείγματος. Εδώ δεν θεωρήσαμε αυτή την δυνατότητα προκειμένου να δώσουμε ξεκάθαρη έμφαση στον αντίκτυπο πολλών βοηθητικών μεταβλητών.

Παρατήρηση 4.3 (Αντίστροφη ερμηνεία του LASSO)

Σχετικά με την πιο πάνω κατασκευή των περιοχών εμπιστοσύνης μπορεί να δοθεί η ακόλουθη αντίστροφη ερμηνεία του LASSO:

$$\hat{\beta}_\alpha = argmin_{\beta \in R^p} E_n\left[(\tilde{y}_{1i} - a\tilde{y}_{2i}) - \tilde{x}'_{ij}\beta\right]^2 + \frac{\lambda}{n}\sum_{j=1}^{p}|\beta_j|\gamma_{aj}, \qquad (4.16)$$

$$\gamma_{aj} = \sqrt{E_n[(\tilde{y}_{1i} - a\tilde{y}_{2i})^2\tilde{x}_{ij}^2]}$$

Εάν $\lambda = 2\Lambda(1-\gamma|W,X)$, τότε το C αντιστοιχεί με την περιοχή $\{\alpha \in R : \hat{\beta}_\alpha = 0\}$. Εάν το $\lambda = 2\Lambda(1-\gamma)$ τότε το C' αντιστοιχεί με την περιοχή $\{\alpha \in R : \hat{\beta}_\alpha = 0\}$. Με λίγα λόγια, για να κατασκευάσουμε τις συγκεκριμένες περιοχές εμπιστοσύνης, συλλέγουμε όλες τις πιθανές τιμές της δομικής παραμέτρου, όπου η πα-

[14] Ονομάζεται η καμπύλη ή η επιφάνεια που περιβάλλει όλες τις καμπύλες ή επιφάνειες τις οποίες παριστάνει μια εξίσωση όταν η παράμετρος που υπάρχει μέσα σε αυτή την εξίσωση παίρνει όλες τις δυνατές τιμές.

λινδρόμηση LASSO της πιθανής δομικής διαταραχής στα εργαλεία παράγει μη-δενικούς συντελεστές στις βοηθητικές μεταβλητές (Chernozhukov και Hansen (2008a), (2008b)).

ΚΕΦΑΛΑΙΟ 5

DANTZIG SELECTOR

5.1 Εισαγωγική ανάλυση του Dantzig-selector

Αναμφισβήτητα, τα τελευταία χρόνια η επιλογή μοντέλων υψηλών διαστάσεων έχει προσελκύσει την προσοχή των επιστημόνων και έχει γίνει το κεντρικό θέμα στο κλάδο της στατιστικής. Ως γνωστόν, η κύρια δυσκολία ενός τέτοιου προβλήματος προέρχεται από την συγγραμικότητα (*collinearity*)[15] μεταξύ των μεταβλητών πρόβλεψης. Είναι σαφές, από γεωμετρικής ερμηνείας ότι αυξάνεται η συγγραμικότητα με την αύξηση της διαστατικότητας.

Στις προηγούμενες ενότητες, εφαρμόζαμε στα αραιά μοντέλα υψηλών διαστάσεων, τις ποινικοποιημένες μεθόδους LASSO, για την αντιμετώπιση του προβλήματος της εμφάνισης πολλών παραμέτρων σε σχέση με τις παρατηρήσεις που είχαμε στη διάθεση μας.

Στο συγκεκριμένο κεφάλαιο θα ασχοληθούμε και πάλι με τις περιπτώσεις των δεδομένων όπου ο αριθμός των παραμέτρων, p, είναι κατά πολύ μεγαλύτερος από το δείγμα, n αλλά θα παρουσιάσουμε μια νέα μέθοδο εκτίμησης.

Υποθέτουμε, ότι έχουμε το μοντέλο:

$$Y = \beta X + z\,, \qquad (5.1)$$

όπου Y είναι οι παρατηρήσεις , $\beta \in R$ είναι ένα διάνυσμα παραμέτρων, X είναι ένας πίνακας με πολύ λιγότερες γραμμές από στήλες ($n \ll p$) και $z \sim N(0, \sigma^2 I_n)$ είναι ένα διάνυσμα ανεξάρτητων τυχαίων κανονικών μεταβλητών.

[15] Το πρόβλημα της συγγραμικότητας παρουσιάζεται όταν μια ανεξάρτητη μεταβλητή συσχετίζεται με μια άλλη ανεξάρτητη μεταβλητή· δηλαδή μέσω της μιας μπορούμε να εκτιμήσουμε τις τιμές της άλλης.

Για τον υπολογισμό του διανύσματος β, θα εισάγουμε ένα νέο εκτιμητή, που προτάθηκε από τους καινοτόμους Candès και Tao (2007). Ο συγκεκριμένος εκτιμητής ονομάζεται Dantzig selector και χρησιμοποιεί την l_1-ελαχιστοποίηση σε συνδυασμό με την κανονικοποίηση των υπολοίπων.

Εξέχουσας σημασίας αποτελεί το γεγονός ότι ο Dantzig selector λύνει ένα γραμμικό πρόβλημα συνήθως πιο γρήγορα και από τις ήδη υπάρχουσες μεθόδους. Να σημειωθεί ότι οι καθηγητές Candès και Tao (2005,2006) και Candès, Romberg και Tao (2006) διαπίστωσαν ότι υπό την Ομοιόμορφη αρχή της αβεβαιότητας (*Uniform Uncertainly Principle, UUP*) και με μεγάλη πιθανότητα, ο Dantzig selector κληρονομεί τον κίνδυνο του εκτιμητή oracle μέχρι ενός λογαριθμικού παράγοντα $logp$ (όπου p είναι ο αριθμός των μεταβλητών) .

5.1.1 Ομοιόμορφη Αρχή της αβεβαιότητας (UUP)

Ένα σημαντικό ερώτημα που γεννάται είναι κατά πόσο είναι εφικτή η εκτίμηση του διανύσματος β μέσα από το y σε υψηλές διαστάσεις, ακόμη και στην «αθόρυβη» (*noiseless*) περίπτωση. Για τον λόγο αυτό κάνουμε χρήση και πάλι της αραιότητας έτσι ώστε να γίνει εφικτή η εύρεση των λύσεων ενός υποκαθορισμένου συστήματος εξισώσεων· το οποίο ορίζεται ως το σύστημα με λιγότερες εξισώσεις παρά αγνώστους . Συγκεκριμένα, υποθέτουμε ότι το διάνυσμα της παραμέτρου β είναι S-αραιό, πράγμα που σημαίνει ότι μόνο οι S από τις p συνιστώσες του διανύσματος β θα είναι μη μηδενικές.

Καταρχήν, ως γνωστόν, το να βρίσκουμε αραιές λύσεις των υποκαθορισμένων συστημάτων των γραμμικών εξισώσεων είναι αρκετά δύσκολο. Αναλυτικότερα έχουμε ότι η αραιή λύση με χρήση της l_0-νόρμας δίνεται από το ακόλουθο πρόβλημα:

$$min_{\beta \in R^p} \|\beta\|_{l_0} \text{ που υπακούει στην εξίσωση } X\beta = y \qquad (5.2)$$

Για την επίλυση αυτού του προβλήματος απαιτούνται διεξοδικές έρευνες για όλα τα υποσύνολα των στηλών του πίνακα X γεγονός που το καθιστά πολύπλοκο. Για τον λόγο αυτό οι ερευνητές αναζήτησαν μια εναλλακτική λύση που

προσεγγίζει παρόμοια το πρόβλημα και γίνεται μέσω της l_1-νόρμας. Συγκεκριμένα, έδειξαν ότι στην «χωρίς θόρυβο» περίπτωση ο υπολογισμός του β γίνεται από την επίλυση του ακόλουθου κυρτού προγράμματος:

$$min_{\tilde{\beta}\in R^p}\left\|\tilde{\beta}\right\|_{l_1} \text{ που υπακούει στην εξίσωση } X\tilde{\beta} = y \qquad (5.3)$$

όπου $\left\|\tilde{\beta}\right\|_{l_1} = \sum_{i=1}^{p}|\tilde{\beta}_i|$ και ο πίνακας $X \in R^{nxp}$ υπακούει στην Ομοιόμορφη αρχή της αβεβαιότητας (*UUP*).

Αντίθετα με την l_0-νόρμα που μετράει της μη μηδενικές συνιστώσες, η l_1- είναι κυρτή. Η l_1-ελαχιστοποίηση βρίσκει χωρίς σφάλμα την περιοχή και το εύρος των μη μηδενικών συνιστωσών του διανύσματος της παραμέτρου β.

Πιο πάνω έγινε αναφορά στην Ομοιόμορφη αρχή της αβεβαιότητας (*UUP*). Ουσιαστικά αυτή η αρχή διατυπώνει ότι για οποιαδήποτε μικρό σύνολο των μεταβλητών πρόβλεψης , αυτά τα n-διανύσματα είναι σχεδόν ορθογώνια μεταξύ τους. Μέσω αυτής της αρχής συνδέεται το γεγονός ότι ο πίνακας σχεδιασμού X υπακούει στην Περιορισμένη ισομετρική υπόθεση (*Restricted Isometry Hypothesis*) για την οποία ακολουθεί παρακάτω μια σύντομη ανάλυση.

5.1.1.1 Περιορισμένες σταθερές ισομετρίας

Ακολουθεί μια περιεκτική αναφορά στις περιορισμένες σταθερές ισομετρίας (*Restricted Isometry Constants*) οι οποίες θα χρησιμοποιηθούν στα Θεωρήματα σχετικά με τον Dantzig selector που θα παρουσιάσουμε στη συνέχεια. Έστω ότι ορίζουμε ο X_T να είναι ο $n \times |T|$ υποπίνακας που δημιουργήθηκε αφαιρώντας τις στήλες του πίνακα σχεδιασμού X που αντιστοιχούν στους δείκτες στο $T \subset \{1,..,p\}$, όπου T ορίζεται ως ο φορέας (*support*) [16]. Έπειτα για κάθε ακέραιο $1 \leq S \leq p$ ορίζουμε την δ_S ως την S-περιορισμένη (S-*restricted*)

[16] Ο φορέας μιας συνάρτησης είναι το σύνολο των σημείων στα οποία η συνάρτηση είναι μη μηδενική, καθώς και το περίβλημα αυτού του συνόλου

σταθερά ισομετρίας του X η οποία είναι η μικρότερη ποσότητα έτσι ώστε ο X_T να υπακούει στην ακόλουθη ανισότητα:

$$(1-\delta_S)\|c\|_{l_2}^2 \leq \|X_T c\|_{l_2}^2 \leq (1+\delta_S)\|c\|_{l_2}^2 \qquad (5.4)$$

για όλα τα υποσύνολα T με $|T| \leq S$ και την ακολουθία των συντελεστών $(c_j)_{\ j\in T}$.

Ουσιαστικά η παραπάνω ιδιότητα αναφέρεται στο γεγονός ότι κάθε ομάδα των στηλών με λιγότερη πληθικότητα από την S-προσέγγιση συμπεριφέρεται σαν ένα ορθοκανονικό σύστημα. Συνάμα, στην περίπτωση που το μέγεθος S υπακούει στη σχέση $\delta_S+\delta_{2S}+\delta_{3S}<1$ τότε η λύση του (5.3) θα είναι οποιοδήποτε αραιό β με το μέγεθος του φορέα T να ικανοποιεί τη σχέση $|T| \leq S$.

Ομοίως ορίζουμε τις $\theta_{S,S'}$ ως τις S,S'-περιορισμένες σταθερές ορθογωνιότητας για $S+S' \leq p$ οι οποίες θα είναι οι μικρότερες ποσότητες έτσι ώστε:

$$|\langle X_T c, X_{T'} c'\rangle| \leq \theta_{S,S'}.\|c\|_{l_2}\|c'\|_{l_2} \qquad (5.5)$$

Το παραπάνω θα ισχύει για τα ξένα σύνολα $T,T' \subseteq \{1,\dots,p\}$ της πληθικότητας, $|T| \leq S$ και $|T'| \leq S'$. Οι μικρές τιμές των περιορισμένων σταθερών ορθογωνιότητας υποδεικνύουν ότι τα ανεξάρτητα υποσύνολα των συμμεταβλητών παράγουν ορθογώνια υποσύνολα.

Επίσης, ισχύει γενικά ότι :

$$\delta_S + \theta_{S,S} + \theta_{S,2S} < 1 \qquad (5.6)$$

όπου για $S' \geq S$ προκύπτει ότι $\delta_{S+S'} - \delta_{S'} \leq +\theta_{S,S'} \leq \delta_{S+S'}$

5.2 Εκτίμηση Dantzig selector

Αυτό που μας προβληματίζει είναι κατά πόσο είναι δυνατό να εκτιμηθεί αξιόπιστα η παράμετρος $\beta \in R^p$ από τα δεδομένα θορύβου (*noise data*) $y \in R^n$ και το

μοντέλο $y = X\beta + z$. Επειδή εκτός του ότι οι παρατηρήσεις του μοντέλου χαρακτηρίζονται από το θόρυβο είναι και πολύ λίγες .

Την λύση στο πρόβλημα έρχεται να δώσει ο Dantzig selector των Emmanuel Candès και Terence Tao (2007). Συγκεκριμένα, η εκτίμηση της παραμέτρου β με δεδομένα που χαρακτηρίζονται από θόρυβο, δίνεται από την λύση του ακόλουθου κυρτού προβλήματος:

$$(DS)\ \ min_{\tilde{\beta}\in R^p}\left\|\tilde{\beta}\right\|_{l_1} \textit{δεδομένου ότι } \|X^*r\|_\infty := sup_{1\le i\le p}|(X^*r)_i| \le \lambda_p.\sigma \quad (5.7)$$

για κάποιο $\lambda_p > 0$, όπου το $r = y - X\tilde{\beta}$ είναι το διάνυσμα των υπολοίπων .

Παρατηρήσεις 5.1 (Dantzig selector και υπόλοιπο r)

Με λίγα λόγια, επιδιώκουμε ένα εκτιμητή με την ελάχιστη πολυπλοκότητα. Στο κλασσικό πρόβλημα της γραμμικής παλινδρόμησης όταν έχουμε $p \le n$ τότε ο εκτιμητής των ελάχιστων τετραγώνων είναι η λύση της κανονικής εξίσωσης $X^T y = X^T X\beta$. Για πραγματικές τιμές ,ο περιορισμός $\|X^*r\|_\infty \le \lambda_p.\sigma$ που ισοδυναμεί με το $\|X^*y - X^*X\beta\|_\infty \le \lambda_p.\sigma$ στο κυρτό πρόγραμμα (*DS*) μπορεί να αντιμετωπισθεί σαν μια χαλάρωση της κανονικής εξίσωσης $X^T y = X^T X\beta$. Ομοίως στην περίπτωση θορύβου η l_1-ελαχιστοποίηση οδηγεί στην αραιότερη λύση στο διάστημα όλων των εφικτών λύσεων.

Επιπλέον, ο περιορισμός σχετικά με το διάνυσμα υπολοίπου r, επιβάλει ότι για κάθε $i \in \{1, \dots, p\}$ ισχύει ότι $|(X^*r)_i| \le \lambda_p.\sigma$ και εγγυάται ότι τα υπόλοιπα θα βρίσκονται εντός του επιπέδου του θορύβου. Η συγκεκριμένη πρόταση γίνεται κατανοητή με την προϋπόθεση ότι οι στήλες του πίνακα σχεδιασμού X έχουν το ίδιο μέγεθος ευκλείδειας και πάντοτε θα υποθέτουμε ότι οι νόρμα τους θα είναι ένα. Τα αποτελέσματα μας θα ισχύουν και για πίνακες με στήλες διαφορετικού μεγέθους και απλά θα πρέπει να αλλάξουμε τη δεξιά πλευρά σε $|(X^*r)_i|$ λιγότερο ή ίσο με $\lambda_p.\sigma$ φορές την ευκλείδια νόρμα της i-οστής στήλης του X, ή σε $|(X^*r)_i| \le \sqrt{1+\delta}\ \lambda_p.\sigma$ δεδομένου ότι όλες οι στήλες έχουν νόρμα μικρότερη του $\sqrt{1+\delta}$.

Όπως φαίνεται οι περιορισμοί του προβλήματος αφορούν το μέγεθος του διανύσματος του σχετικού υπολοίπου X^*r και όχι το μέγεθος του διανύσματος του υπολοίπου r. Αυτό δεν είναι καθόλου τυχαίο μιας και υπάρχουν συγκεκριμένοι λόγοι για αυτή την κίνηση. Καταρχήν, έστω ότι έχουμε ένα ορθοκανονικό μετασχηματισμό που εφαρμόζεται στα δεδομένα, δίνοντας $y' = Uy$, όπου U^*U είναι ο ταυτοτικός. Είναι ξεκάθαρο ότι για να είναι καλή μια διαδικασία εκτίμησης για το β δεν θα πρέπει να εξαρτάται από το U. Ενώ αν θέλουμε να επιστρέψουμε στο αρχικό πρόβλημα απλά εφαρμόζουμε το U^*. Επιπλέον, να σημειωθεί ότι η εκτίμηση με την βοήθεια του Dantzig selector (5.7) είναι στην πραγματικότητα αναλλοίωτη σε σχέση με τον ορθοκανονικό μετασχηματισμό που εφαρμόζεται στο διάνυσμα των δεδομένων, δεδομένου ότι η εφικτή περιοχή είναι αναλλοίωτη : $(UX)^T\left(UX\,\tilde{\beta} - Uy\right) = X^*\left(X\,\tilde{\beta} - y\right)$. Αντίθετα, αν είχαμε ορίσει την εφικτή περιοχή με το $sup_i|r_i|$ να είναι μικρότερο από το προσαρμοσμένο όριο (*fixed threshold*), τότε η διαδικασία εκτίμησης σε αυτή την περίπτωση δεν θα ήταν αναλλοίωτη.

Ένας άλλος λόγος είναι ότι αναμφισβήτητα επιθυμούμε να συμπεριλάβουμε στο μοντέλο επεξηγηματικές μεταβλητές που θα είναι ιδιαίτερα συσχετισμένες με τα δεδομένα y. Έστω, ότι εξετάζουμε την κατάσταση κατά την οποία το διάνυσμα υπολοίπου r είναι ίσο με την στήλη X^i του πίνακα σχεδιασμού X. Σε αυτή την περίπτωση, υποθέτουμε για λόγους απλότητας ότι όλα τα στοιχεία της στήλης X^i έχουν περίπου το ίδιο μέγεθος, το οποίο είναι περίπου $1/\sqrt{n}$. Συγχρόνως υποθέτουμε ότι το σ είναι ελαφρώς μεγαλύτερο από $1/\sqrt{n}$. Στη περίπτωση που χρησιμοποιούσαμε ένα περιορισμό της μορφής $sup_i|r_i| \leq \lambda_n\sigma$, τότε το διάνυσμα υπολοίπων θα ήταν εφικτό, πράγμα που δεν θα είχε κάποιο νόημα. Αντίθετα, ένα τέτοιο διάνυσμα υπολοίπων δεν θα ήταν εφικτό για τον εκτιμητή Dantzig selector για λογικές τιμές του επιπέδου του θορύβου , και η i-οστή τιμή θα συμπεριλαμβανόταν στο μοντέλο.

Ακολούθως ως γνωστόν, το πρόγραμμα του Dantzig selector είναι κυρτό και μπορεί εύκολα να ξαναδοκιμαστεί σαν ένα γραμμικό πρόγραμμα,

$$\begin{gathered} min \sum_i u_i \text{ δεδομένου ότι } u \leq \tilde{\beta} \leq u \text{ και} \\ -\lambda_p\sigma\mathbf{1} \leq X^*(y - X\tilde{\beta}) \leq \lambda_p\sigma\mathbf{1} \end{gathered} \tag{5.8}$$

όπου οι μεταβλητές βελτιστοποίησης είναι οι u , $\tilde{\beta} \in R^p$ και $\mathbf{1}$ είναι ένα μοναδιαίο διάνυσμα p-διαστάσεων. Με τον τρόπο αυτό, η διαδικασία εκτίμησης γίνεται υπολογιστικά ελέγξιμη. Πράγματι, σύμφωνα με τους Boyd και Vandenberghe (2004), υπάρχει μια μεγάλη οικογένεια αλγορίθμων για την επίλυση τέτοιων προβλημάτων.

Ένα από τα σημαντικότερα αποτελέσματα αυτής της ενότητας είναι το ακόλουθο θεώρημα, το οποίο με λίγα λόγια διατυπώνει την ακρίβεια του Dantzig selector.

Θεώρημα 5.1

Υποθέτουμε ότι $\beta \in R^p$ είναι οποιοδήποτε S-αραιό διάνυσμα των παραμέτρων υπακούοντας στη σχέση $\delta_{2S} + \theta_{S,2S} < 1$. Επιλέγουμε $\lambda_p = \sqrt{2logp}$ για τον περιορισμό του διανύσματος των υπολοίπων στη σχέση (5.7).Τότε, με μεγάλη πιθανότητα, το $\hat{\beta}$ υπακούει στην ακόλουθη ανισότητα:

$$\left\|\hat{\beta} - \beta\right\|_{l_2}^2 \leq C_1^2 \cdot (2logp) \cdot S \cdot \sigma^2 \quad \text{με } C_1 = 4/(1 - \delta_S - \theta_{S,2S}) \qquad (5.9)$$

Άρα για μικρές τιμές της ποσότητας $\delta_S + \theta_{S,2S}$ θα έχουμε ότι $C_1 \approx 4$**.** Αν επιλέξουμε $\lambda_p := \sqrt{2(1+a)logp}$ για κάθε $a \geq 0$, το φράγμα θα ισχύει με πιθανότητα που θα ξεπερνά το $1 - \left(\sqrt{\pi logp}.p^a\right)^{-1}$ με τον όρο ότι το λ_p^2 θα αντικαθιστά την ποσότητα $2logp$ στην σχέση (5.9).

■

Παρατηρήσεις 5.2

Αυτό που κάνει εντύπωση από το παραπάνω θεώρημα είναι το γεγονός ότι όχι μόνο πλέον μπορούμε να εκτιμήσουμε αξιόπιστα το διάνυσμα των παραμέτρων β μέσα από τις λίγες παρατηρήσεις που έχουμε, αλλά το μέσο τετραγωνικό σφάλμα είναι ανάλογο με τον πραγματικό αριθμό των αγνώστων επί το επίπεδο

του θορύβου σ^2. Το αξιοσημείωτο σε αυτό είναι το γεγονός ότι μπορεί να επιτευχθεί το συγκεκριμένο κατόρθωμα με την λύση ενός απλού γραμμικού μοντέλου.

Επιπλέον και αγνοώντας τον λογαριθμικό παράγοντα, η στατιστική λογική μας λέει ότι η σχέση (5.9) δεν επιδέχεται βελτίωσης. Για να το δικαιολογήσουμε αυτό, υποθέτουμε ότι έχουμε διαθέσιμο ένα oracle αφήνοντας μας να γνωρίζουμε εκ των προτέρων τη θέση των S μη μηδενικών στοιχείων του διανύσματος της παραμέτρου, δηλαδή $T_0 \coloneqq \{i: \beta_i \neq 0\}$.Με λίγα λόγια θα γνωρίζουμε το σωστό μοντέλο νωρίτερα.

Η συγκεκριμένη πληροφορία είναι χρήσιμη για να κατασκευαστεί ένας ιδανικός εκτιμητής β^* χρησιμοποιώντας την προβολή των ελάχιστων τετραγώνων:

$$\beta^*_{T_0} = (X^T_{T_0} X_{T_0})^{-1} X^T_{T_0} y \qquad (5.10)$$

όπου το $\beta^*_{T_0}$ είναι ο περιορισμός του β^* στο σύνολο T_0 και να τεθεί το b με μηδέν εκτός του T.

Επομένως , θα έχουμε ότι:

$$\begin{gathered} \beta^* = \beta + (X^T_{T_0} X_{T_0})^{-1} X^T_{T_0} z \\ E\|\beta^* - \beta\|^2_{l_2} = E\left\|(X^T_{T_0} X_{T_0})^{-1} X^T_{T_0} z\right\|^2_{l_2} = \sigma^2 Tr((X^T_{T_0} X_{T_0})^{-1}) \end{gathered} \qquad (5.11)$$

Πλέον, όλες οι ιδιοτιμές του $X^T_{T_0} X_{T_0}$ θα ανήκουν στο διάστημα $[1-\delta_s, 1+\delta_s]$ και το ιδανικό αναμενόμενο μέσο τετραγωνικό σφάλμα θα υπακούει στην ακόλουθη ανισότητα:

$$E\|\beta^* - \beta\|^2_{l_2} \geq \frac{1}{1+\delta_s} \cdot S \cdot \sigma^2 \qquad (5.12)$$

Επίσης, παρατηρούμε ότι γίνεται μια επιλογή του $\lambda_p = \sqrt{2logp}$ όπου στην περίπτωση ορθογώνιου σχεδιασμού αυτό ισοδυναμεί με $\sqrt{2logn}$.

Το Θεώρημα 5.1 καταλήγει στο γεγονός ότι ο ελάχιστος l_1 −εκτιμητής επιτυγχάνει μια απώλεια εντός ενός λογαριθμικού παράγοντα του ιδανικού μέσου τετραγωνικού σφάλματος. Ο λογαριθμικός αυτός παράγοντας είναι στην ουσία το τίμημα που πληρώνουμε εξαιτίας της αναπροσαρμοστικότητας, δηλαδή να μην γνωρίζουμε από πριν που είναι στην πραγματικότητα οι μη μηδενικές τιμές της παραμέτρου.

Με λίγα λόγια , η διαδικασία ανάκτησης αν και είναι εξαιρετικά μη γραμμική, είναι ευσταθής υπό την παρουσία του θορύβου. Αυτό είναι ιδιαίτερα ενδιαφέρον μιας και ο πίνακας X του γραμμικού μοντέλου $y = X\beta + z$ είναι ορθογώνιος, έχοντας δηλαδή πολύ περισσότερες στήλες σε σχέση με τις γραμμές. Ως εκ τούτου, οι περισσότερες από τις ιδιάζουσες τιμές (*singular values*) [17] του είναι μηδέν.

Με την επίλυση του Dantzig selector, προσπαθούμε ουσιαστικά να αντιστρέψουμε τη δράση του πίνακα σχεδιασμού X στο κρυμμένο μας β υπό την παρουσία θορύβου. Το γεγονός ότι αυτή η διαδικασία αντιστροφής του πίνακα κρατάει την διαταραχή από το να τείνει στο άπειρο είναι ίσως κάτι που δεν περιμέναμε εξ' αρχής.

Προφανώς το αποτέλεσμα μας θα ήταν ιδιαίτερα ενδιαφέρον εάν μπορούσαμε να εκτιμήσουμε την τάξη των n παραμέτρων με n παρατηρήσεις. Δηλαδή , θα επιθυμούσαμε να ισχύει η συνθήκη $\delta_{2S} + \theta_{S,2S} < 1$ για μεγάλο αριθμό τιμών της $S\cdot$ παραδείγματος χάριν όσο πιο κοντά γίνεται στο n. Να σημειωθεί ότι για $2S > n$, $\delta_{2S} \geq 1$ εφόσον κάθε πίνακας με περισσότερες από n στήλες πρέπει να είναι ιδιάζων, πράγμα που σημαίνει ότι $S < n/2$.

Γενικά έχει αποδειχθεί ότι όσον αφορά τους πίνακες σχεδιασμούς X, η παραπάνω συνθήκη ισχύει για πολύ σημαντικές τιμές του S. Αναλυτικότερα, υποθέτουμε ότι έχουμε ένα τυχαίο πίνακα X με στοιχεία που ακολουθούν την γκαουσιανή κατανομή. Τότε με μεγάλη πιθανότητα, η συνθήκη ισχύει για

[17] Έστω ότι έχουμε την ανάλυση του mxm πίνακα $M = U\Sigma V^*$. Όπου Σ είναι ένας mxn ορθογώνιος διαγώνιος πίνακας με μη αρνητικές τιμές στην διαγώνιο. Τότε τα διαγώνια στοιχεία $\Sigma_{i,i}$ του Σ είναι γνωστά ως οι ιδιάζουσες τιμές του M.

$S = O(n/\log(p/n))$. Δηλαδή, αυτή η ρύθμιση απαιτεί μόνο $O(log\left(\frac{p}{n}\right))$ παρατηρήσεις για κάθε μη μηδενική τιμή της παραμέτρου.

5.3 Ανισότητες oracle

Είναι αδιαμφησβήτητο ότι το Θεώρημα 5.1 μας δίνει σημαντικά αποτελέσματα που σχετίζονται με τον Dantzig selector · εντούτοις υπάρχουν και περιπτώσεις που τα πράγματα μπορούν να απλοποιηθούν αρκετά.

Σε αυτή την περίπτωση, έστω ότι επιλέγουμε το β να είναι πολύ μικρό έτσι ώστε να είναι μικρότερο από το επίπεδο θορύβου (*noise level*) και να ισχύει δηλαδή ότι $|\beta_\iota| \ll \sigma$ για κάθε i. Κάνοντας αυτή την υπόθεση, είναι δυνατό να θέσουμε τον εκτιμητή με μηδέν, δηλαδή $\hat{\beta} = 0$. Η απώλεια του τετραγωνικού σφάλματος τότε θα ήταν $\sum_{i=1}^{p}|\beta_\iota|^2$, η οποία ενδεχομένως να είναι πολύ μικρότερη από το σ^2 επί τον αριθμό των μη μηδενικών στοιχείων του β. Κατά κάποιο τρόπο , αυτή είναι μια κατάσταση κατά την οποια το τετράγωνο της μεροληψίας θα είναι πολύ μικρότερο από την διακύμανση.

Ιδανικό μέσο τετραγωνικό σφάλμα (MSE)

Μια εναλλακτική πρόταση θα ήταν να ζητηθεί μια σχεδόν βέλτιστη εξισορρόπηση συντεταγμένης με συντεταγμένη (*near optimal trade-off coordinate by coordinate*). Αναλυτικότερα, έστω ότι για λόγους απλότητας ο πίνακας σχεδιασμού X αντιστοιχεί στον ταυτοτικό πίνακα και επίσης ισχύει $y{\sim}N(\beta, \sigma^2 I_p)$. Τότε υποθέτουμε ότι έχουμε στην διάθεση μας ένα εκτιμητή oracle που μας επιτρέπει να γνωρίζουμε εκ των προτέρων ποιες συντεταγμένες του β είναι σημαντικές, με άλλα λόγια το σύνολο των δεικτών για τους οποίους ισχύει ότι $|\beta_i| > \sigma$. Έπειτα, έχοντας ως εφόδιο τον oracle , θέτουμε όπου $\beta_i^* = y_i$ για κάθε δείκτη που ανήκει στο σύνολο των σημαντικών και για τους υπόλοιπους δείκτες θέτουμε $\beta_i^* = 0$.

Τότε, το αναμενόμενο μέσο τετραγωνικό σφάλμα του ιδανικού αυτού εκτιμητή θα είναι:

$$E\|\beta^* - \beta\|_{l_2}^2 = \sum_{i=1}^{p} \min(\beta_i^2, \sigma^2) \qquad (5.13)$$

Από εδώ και πέρα θα αναφερόμαστε στο πιο πάνω τύπο (5.13) ως το ιδανικό μέσο τετραγωνικό σφάλμα (*Mean Square Error, MSE*). Σύμφωνα με τους Donoho και Johnstone (1994), με οριακό επίπεδο (*threshold level*) περίπου $\sqrt{2logp} \cdot \sigma$, επιτυγχάνεται το ιδανικό MSE σε ένα λογαριθμικό παράγοντα ανάλογο του $logp$.

Σχετικά με το γραμμικό μοντέλο θεωρούμε τον εκτιμητή ελάχιστων τετραγώνων ως $\hat{\beta}_I = (X_I^T X_I)^{-1} X_I^T y$ και τον ιδανικό εκτιμητή ελάχιστων τετραγώνων ο οποίος ελαχιστοποιεί το αναμενόμενο μέσο τετραγωνικό σφάλμα ως

$$\beta^* = \ argmin_{I \subset \{1,\dots,p\}} E\left\|\beta - \hat{\beta}_I\right\|_{l_2}^2 \qquad (5.14)$$

Δηλαδή, θα εφαρμόζεται ο ιδανικός εκτιμητής στα μοντέλα ελάχιστων τετραγώνων και με βάση τον oracle θα μας υποδεικνύει πιο μοντέλο να επιλέξουμε. Ονομάζεται ιδανικός επειδή μπορούμε να μην υπολογίσουμε την ποσότητα $E\left\|\beta - \hat{\beta}_I\right\|_{l_2}^2$ μιας και δεν γνωρίζουμε το β · αν και στη συνέχεια θα προσπαθήσουμε να το εκτιμήσουμε. Αυτό θα μπορούσε να θεωρηθεί ως ένα σημείο αναφοράς έτσι ώστε να διερωτηθούμε κατά πόσο ένας πραγματικός εκτιμητής θα μπορούσε να υπακούει στην ακόλουθη σχέση με μεγάλη πιθανότητα:

$$\left\|\hat{\beta} - \beta\right\|_{l_2}^2 = O(logp) \cdot E\|\beta - \beta^*\|_{l_2}^2 \qquad (5.15)$$

Κατά κάποιο τρόπο η σχέση (5.13) αποτελεί ένα υποκατάστατο για τον ιδανικό κίνδυνο (*ideal risk*) $E\|\beta - \beta^*\|_{l_2}^2$. Πράγματι, έστω ότι το I είναι ένα προσαρμοσμένο υποσύνολο των δεικτών και θεωρούμε ότι y είναι η παλινδρόμηση σε αυτό το υποσύνολο και όπου β_I είναι ο περιορισμός του β μέσα στο υποσύνολο I. Τότε το σφάλμα του εκτιμητή $\hat{\beta}_I$ θα δίνεται από την ακόλουθη σχέση:

$$\left\|\hat{\beta}_I - \beta\right\|_{l_2}^2 = \left\|\hat{\beta}_I - \beta_I\right\|_{l_2}^2 + \|\beta_I - \beta\|_{l_2}^2 \qquad (5.16)$$

Αναλυτικότερα, ο πρώτος όρος θα είναι ίσος με:

$$\hat{\beta}_I - \beta_I = (X_I^T X_I)^{-1} X_I^T X \beta_{I^c} + (X_I^T X_I)^{-1} X_I^T z \qquad (5.17)$$

Ενώ το αναμενόμενο μέσο τετραγωνικό σφάλμα θα δίνεται από την ακόλουθη σχέση:

$$E\left\|\hat{\beta}_I - \beta_I\right\|^2 = \|(X_I^T X_I)^{-1} X_I^T X \beta_{I^c}\|_{l_2}^2 + \sigma^2 Tr((X_I^T X_I)^{-1}) \qquad (5.18)$$

Έτσι, αυτός ο όρος θα υπακούει στην ακόλουθη ανισότητα:

$$E\left\|\hat{\beta}_I - \beta_I\right\|^2 \geq \frac{1}{1+\delta_{|I|}} \cdot |I| \cdot \sigma^2 \qquad (5.19)$$

Για όλα τα σύνολα I με $|I| \leq S$ και $\delta_S < 1$, θα έχουμε :

$$E\left\|\hat{\beta}_I - \beta\right\|^2 \geq \frac{1}{2} \cdot \left(\sum_{i \in I^c} \beta_i^2 + |I| \cdot \sigma^2\right) \qquad (5.20)$$

που μας δίνει το ιδανικό μέσο τετραγωνικό σφάλμα το οποίο είναι φραγμένο όπως φαίνεται παρακάτω:

$$\|\beta^* - \beta\|_{l_2}^2 \geq \frac{1}{2} \cdot min\left(\sum_{i \in I^c} \beta_i^2 + |I| \cdot \sigma^2\right) = \frac{1}{2} \cdot \sum_i \min(\beta_i^2, \sigma^2) \qquad (5.21)$$

Με άλλα λόγια ο ιδανικός κίνδυνος, είναι κάτω φραγμένος από την σχέση (5.13). Όπως έχουμε δει η σχέση αυτή είναι ουσιώδες μιας και έχει φυσική ερμηνεία στους όρους της ιδανικής μεροληψίας και διασποράς, δηλαδή

$$\sum_i min(\beta_i^2, \sigma^2) = min_{I \subset \{1,\dots,p\}} \|\beta - \beta_I\|_{l_2}^2 + |I| \cdot \sigma^2 \qquad (5.22)$$

Μετά από όλα αυτά καταλήγουμε σε ένα σημαντικό συμπέρασμα. Ότι με βάση τα δεδομένα που δίνονται από την σχέση $Y = \beta X + z$ και χωρίς να γνωρίζουμε τις σημαντικές συντεταγμένες του β καθώς και χωρίς να είμαστε σε θέση να παρατηρούμε κατευθείαν τις τιμές της παραμέτρου, ένας εκτιμητής ο οποίος μπορεί να πλησιάζει πολύ στο αποτέλεσμα της σχέσης (5.15) είναι ο Dantzig selector.

ΘΕΩΡΗΜΑ 5.2

Επιλέγουμε ένα $t > 0$ και ορίζουμε $\lambda_p := (1 + t^{-1})\sqrt{2logp}$ στη σχέση (5.7) που μας δίνει τον Dantzig selector. Τότε εάν το β είναι S-αραιό με $\delta_{2S} + \theta_{S,2S} < 1 - t$, ο εκτιμητής μας θα υπακούει στη σχέση:

$$\left\|\hat{\beta} - \beta\right\|_{l_2}^2 \leq C_2^2 \cdot \lambda_p^2 \cdot \left(\sigma^2 + \sum_{i=1}^p \min(\beta_i^2, \sigma^2)\right) \qquad (5.23)$$

με μεγάλη πιθανότητα (την πιθανότητα που είχαμε και πριν για $\lambda_p := \left(\sqrt{1+\alpha} + t^{-1}\right) \times \sqrt{2logp}$). Η σταθερά C_2 εξαρτάται μόνο από τα μεγέθη δ_{2S} και $\theta_{S,2S}$.

■

Παρατηρήσεις 5.3

Σε αυτό το σημείο τονίζεται ότι η ανισότητα (5.23) είναι μη ασυμπτωτική και η ανάλυση μας δίνει στην πραγματικότητα τις σταθερές σε λυμένη μορφή. Συγκεκριμένα, αποδείξαμε ότι :

$$C_2 = 2\frac{C_0}{1-\delta-\theta} + 2\frac{\theta(1+\delta)}{(1-\delta-\theta)^2} + \frac{1+\delta}{1-\delta-\theta} \qquad (5.24)$$

με $C_0 \coloneqq 2\sqrt{2}\left(1+\frac{1-\delta^2}{1-\delta-\theta}\right)+(1+1/\sqrt{2})\frac{(1+\delta)^2}{1-\delta-\theta}$.

Και στις δυο περιπτώσεις θέτουμε για ευκολία ότι $\delta \coloneqq \delta_{2S}$ και $\theta = \theta_{S,2S}$. Για δ και θ να είναι μικρά , τότε η σταθερά C_2 προσεγγίζεται ως:

$$C_2 \approx 2\left(4\sqrt{2}+1+\frac{1}{\sqrt{2}}\right)+1 \leq 16 \qquad (5.25)$$

Η συνθήκη που επιβάλει ότι δ_{2S} + $\theta_{S,2S} < 1$ (ή μικρότερο από το $1-t$) έχει μια φυσική ερμηνεία στη αναγνωρισιμότητα των όρων του μοντέλου. Έστω ότι έχουμε ένα υποπίνακα ελλιπής μορφής $X_{T\cup T'}$ με $2S$ στήλες και μικρότερη ιδιοτιμή να είναι $0 = 1-\delta_{2S}$, και με τους δείκτες T και T' όπου ο καθένας είναι μεγέθους S. Τότε θα υπάρχει ένα διάνυσμα h που θα υπακούει στη σχέση $Xh = 0$ το οποίο αναλύεται ως $h = \beta - \beta'$, όπου το β έχει φορέα το T. Αντίστοιχα ισχύουν και για το β' και προκύπτει ότι $X\beta = X\beta'$.

Εν ολίγοις, το παραπάνω υποστηρίζει ότι το μοντέλο δεν θα είναι προσδιορίσιμο αφού τα β και β' είναι S -αραιά. Με άλλα λόγια χρειαζόμαστε $\delta_{2S} < 1$. Η απαίτηση δ_{2S} + $\theta_{S,2S} < 1$ είναι πολύ ισχυρή αφού θέτει ένα χαμηλότερο φράγμα στις ιδιάζων τιμές των υποπινάκων και στην ουσία εμποδίζει καταστάσεις στις οποίες θα μπορούσε να υπάρχει πολυσυγγραμικότητα μεταξύ υποσυνόλων των αλγορίθμων πρόβλεψης.

5.4 Ιδανική επιλογή μοντέλου μέσω γραμμικού προγραμματισμού

Η διαδικασία εκτίμησης είναι μια έμμεση μέθοδος για την επιλογή ενός επιθυμητού υποσυνόλου των αλγορίθμων (μέσων) πρόβλεψης που βασίζονται στα δεδομένα θορύβου $y = X\beta + z$, μεταξύ όλων των υποσυνόλων των μεταβλητών. Είναι σημαντικό να τονισθεί ότι στη μέχρι τώρα ανάλυση δεν υπάρχει κάπου η αναφορά ότι ο αριθμός των παραμέτρων επιβάλλεται να είναι μεγαλύτερος από το μέγεθος του δείγματος · γεγονός που δείχνει ότι ο Dantzig selector μπορεί να χαρακτηρισθεί ως μια γενικότερη επιλογή μεταβλητής.

Στην πάροδο των χρόνων προτάθηκαν αρκετές διαδικασίες που αφορούν την επιλογή μοντέλου με την πιο συνηθισμένη να είναι η διαδικασία κανονικής επιλογής (*canonical selection procedure*), των Foster και George (1994), που ορίζεται παρακάτω:

$$argmin_{\tilde{\beta}\in R^p}\left\|y - X\tilde{\beta}\right\|_{l_2}^2 + \Lambda \cdot \sigma^2 \cdot \left\|\tilde{\beta}\right\|_{l_0},\ \left\|\tilde{\beta}\right\|_{l_0} \coloneqq |\{i: \tilde{\beta}_i \neq 0\}| \qquad (5.26)$$

Δημοφιλές διαδικασίες επιλογής όπως είναι η AIC, C_p, BIC και RIC είναι της ίδιας μορφής απλά με διαφορετικές τιμές της παραμέτρου Λ. Στο τομέα της διαδικασίας επιλογής υπάρχουν ουσιαστικά δύο κρίσιμα προβλήματα. Καταρχήν, η εύρεση του ελαχίστου της σχέσης (5.26) είναι τύπου NP-hard πρόβλημα (Natarajan (1995)). Το να λύνεις τέτοια προβλήματα απαιτεί να κάνεις εξαντλητικές έρευνες πάνω σε όλα τα υποσύνολα των στηλών του X, μια διαδικασία η οποία είναι εκ φύσεως συνδυαστική και έχει εκθετική πολυπλοκότητα μιας και θα υπάρχουν 2^p τέτοια υποσύνολα. Βέβαια, υπάρχουν και περιπτώσεις όπου η λύση είναι πιο εφικτή · παραδείγματος χάριν όταν ο πίνακας X είναι ορθοκανονικός. Με άλλα λόγια η επίλυση ενός προβλήματος για την επιλογή μοντέλου είναι εφικτή μόνο όταν το p δεν είναι πολύ μεγάλο.

Έπειτα είναι σημαντικό να τονισθεί ότι η εκτίμηση του β και του $X\beta$ είναι δύο εντελώς διαφορετικά προβλήματα· ειδικότερα όταν το p είναι πολύ μεγαλύτερο του n. Τα τελευταία χρόνια οι επιστήμονες ανακάλυψαν εναλλακτικές λύσεις για να ξεπεραστούν οι παραπάνω δυσκολίες. Μια εκ αυτών είναι η LASSO που αναλύθηκε σε προηγούμενη ενότητα ,γνωστή και ως Basis Pursuit, η οποία χαλαρώνει τη μηδενική νόρμα $\left\|\tilde{\beta}\right\|_{l_0}$ μέσω της κυρτής l_1-νόρμας $\left\|\tilde{\beta}\right\|_{l_1}$.

5.5 Επέκταση στις σχεδόν αραιές παραμέτρους

Μέχρι στιγμής, έχουμε εξετάσει την εκτίμηση διανυσμάτων αραιών παραμέτρων, δηλαδή με ένα αριθμό S μη μηδενικών στοιχείων που υπακούνε στη σχέση $\delta_{2S} + \theta_{S,2S} < 1$. Έχουμε ήδη εξηγήσει ότι η συνθήκη αυτή είναι αναγκαία κατά κάποιο τρόπο αλλιώς θα είχαμε ένα πρόβλημα «aliasing», δηλαδή μια

κατάσταση κατά την οποία θα ίσχυε ότι $X\beta = X\beta'$ παρόλο που τα β και β' πιθανό να είναι εντελώς διαφορετικά. Ωστόσο, με την επέκταση μας σε αποτελέσματα που αφορούν μη αραιά αντικείμενα ο ένας τύπος περιορισμού πιθανό να επιβάλλει στους άλλους να αφαιρέσουν την πιθανότητα μια ισχυρής "aliasing". Μπορεί να υπάρχουν πολλοί τέτοιοι περιορισμοί και εξετάζουμε ένα από αυτούς που αφαιρεί ένα μέρος της συνθήκης των στοιχείων του β.

Αναδιατάσσουμε τα στοιχεία του β κατά φθίνουσα τάξη μεγέθους, δηλαδή $|\beta_{(1)}| \geq |\beta_{(2)}| \geq \ldots \left|\beta_{(p)}\right|$ και υποθέτουμε ότι το k-οστό μεγαλύτερο στοιχείο θα υπακούει στη σχέση

$$\left|\beta_{(k)}\right| \leq R \cdot k^{-1/s} \qquad (5.27)$$

για κάποιο θετικό R και για $s \leq 1$.

Αυτό που μας απασχολεί τώρα είναι κατά πόσο ο εκτιμητής μας θα μπορεί να επιτύχει ένα σφάλμα πολύ κοντά σε αυτό που δίνεται από την σχέση (5.13). Καταρχήν, για να το επιτύχουμε αυτό θα πρέπει να είμαστε σε θέση να εκτιμήσουμε αξιόπιστα όλες τις συντεταγμένες που είναι σημαντικά μεγαλύτερες από το επίπεδο θορύβου, δηλαδή να ισχύει $|\beta_i| \geq \sigma$. Έστω ότι θέτουμε $S = |\{i: |\beta_i| > \sigma\}|$. Τότε στην περίπτωση που ισχύει $\delta_{2S} + \theta_{S,2S} < 1$ αυτό θα μπορούσε να επιτευχθεί αλλά πιθανό να μην έχουμε αρκετές παρατηρήσεις για να είμαστε σε θέση να εκτιμήσουμε τις πολλές σταθερές που έχουμε.

Επιπλέον σχετικά με το ερώτημα που τέθηκε, μια άλλη παρατήρηση που θα σημειώσουμε αφορά το $\beta \in R^p$ που υπακούει στην προηγούμενη σχέση (5.27) και θα έχουμε

$$\sum_i \min(\beta_i^2, \sigma^2) = S \cdot \sigma^2 + \sum_{i \geq S+1} \left|\beta_{(i)}\right|^2 \leq C \cdot (S \cdot \sigma^2 + R^2 S^{-2r}) \qquad (5.28)$$

όπου $r = \frac{1}{s} - \frac{1}{2}$.

Με βάση αυτού του πορίσματος εξάγεται το ακόλουθο Θεώρημα.

Θεώρημα 5.3

Υποθέτουμε ότι το $\beta \in R^p$ υπακούει στη σχέση $|\beta_{(k)}| \leq R \cdot k^{-1/s}$ και ορίζουμε το S_* να είναι σταθερό έτσι ώστε $\delta_{2S_*} + \theta_{S_*,2S_*} < 1$. Επιλέγουμε το λ_p όπως δίνεται από το Θεώρημα 5.1, δηλαδή $\lambda_p = \sqrt{2logp}$. Τότε το $\hat{\beta}$ υπακούει στην ακόλουθη σχέση με μεγάλη πιθανότητα

$$\left\|\hat{\beta} - \beta\right\|_{l_2}^2 \leq min_{1 \leq S \leq S^*} C_3 \cdot 2logp \cdot (S \cdot \sigma^2 + R^2 S^{-2r}) \qquad (5.29)$$

Να σημειωθεί ότι για κάθε β που υπακούει στη σχέση (5.27) τότε $|\{i: |\beta_i| > \sigma\}| \leq \left(\frac{R}{\sigma}\right)^{\frac{1}{s}}$. Αν έχουμε ότι $S_* \geq \left(\frac{R}{\sigma}\right)^{\frac{1}{s}}$ τότε προκύπτει ότι η σχέση (5.29) παίρνει την ακόλουθη μορφή:

$$\left\|\hat{\beta} - \beta\right\|_{l_2}^2 \leq O(logp) \cdot R^{2/(2r+1)} \cdot (\sigma^2)^{2r/(2r+1)} \qquad (5.30)$$

Η σχέση αυτή αποτελεί τον γνωστό ελαχιστομέγιστο (*minimax*) ρυθμό για τις κατηγορίες των αντικειμένων που παρουσιάζουν ένα είδος μείωσης. Παρόλο που έχουμε ότι $n \ll p$, ο Dantzig selector ανακτά τον ελαχιστομέγιστο ρυθμό εάν είμαστε σε θέση να μετρήσουμε όλες τις συντεταγμένες του β ευθέως μέσω της κατανομής $\tilde{y} \sim N(\beta, \sigma^2 I_p)$. Στην περίπτωση που ισχύει ότι $S_* \leq \left(\frac{R}{\sigma}\right)^{\frac{1}{s}}$, η μέθοδος παθαίνει κορεσμό γιατί δεν υπάρχουν αρκετά δεδομένα για να ανακτήσει τον ελαχιστομέγιστο ρυθμό, και μπορεί να εγγυηθεί μόνο μια τετραγωνική απώλεια σχετικά με το $O(logp)(R^2 {S_*}^{-2r} + S_* \cdot \sigma^2)$. Τέλος, να σημειωθεί ότι αυτό το σφάλμα είναι ελεγχόμενο.

5.6 Gauss-Dantzig selector

Όταν ο πίνακας σχεδιασμού X είναι ορθογώνιος, τότε ο Dantzig selector $\hat{\beta}$ είναι

ο l_1-ελαχιστοποιητής του περιορισμού $\left\|X^*y - \hat{\beta}\right\|_{l_\infty} \leq \lambda_p \cdot \sigma$. Απ' αυτό συνεπάγεται ότι ο $\hat{\beta}$ είναι απλά η ήπια οριακή (*soft-thresholded*) εκδοχή του X^*y στο επίπεδο $\lambda_p \cdot \sigma$, και έτσι

$$\hat{\beta}_i = max\big(|(X^*y)_i| - \lambda_p \cdot \sigma, 0\big) sgn((X^*y)_i) \qquad (5.31)$$

Με λίγα λόγια το X^*y είναι μετατοπισμένο προς την αρχή. Υπάρχουν πολλές απλές μέθοδοι για την διόρθωση της μεροληψίας και για την αύξηση της απόδοσης. Θα εξετάσουμε μια εκ αυτών η οποία βασίζεται σε δύο στάδια .

Αρχικά εκτιμάμε το $I = \{i\colon \beta_i \neq 0\}$ μέσω του $\hat{I} = \{i\colon \hat{\beta}_i \neq 0\}$ και το β όπως δίνεται από την σχέση (5.7). Γενικότερα θα μπορούσαμε να πούμε ότι η εκτίμηση γίνεται με το $\hat{I} = \{i\colon \hat{\beta}_i > \alpha \cdot \sigma\}$ για $\alpha \geq 0$. Ακολούθως, κατασκευάζουμε τον εκτιμητή $\hat{\beta}_{\hat{I}} = \left(X_{\hat{I}}^T X_{\hat{I}}\right)^{-1} X_{\hat{I}}^T y$ και θέτουμε τις υπόλοιπες συνιστώσες με μηδέν.

Ως εκ τούτου βασιζόμαστε στον Dantzig selector για να εκτιμήσουμε το μοντέλο I, και κατασκευάζουμε ένα νέο εκτιμητή με παλινδρόμηση των δεδομένων y στο μοντέλο $\hat{I}$. Θα αναφερόμαστε σε αυτή την παραλλαγή ως τον Gauss-Dantzig selector. Αναμένουμε ότι τα Θεωρήματα που αναλύθηκαν προηγουμένως, σχετικά με τον απλό Dantzig selector, θα ισχύουν και για αυτή την παραλλαγή.

Αν και αποδείξαμε τα αποτελέσματα μας στην περίπτωση όπου το z είναι ένα διάνυσμα ανεξάρτητων και ταυτοτικά κατανεμημένων γκαουσιανών μεταβλητών, είναι βέβαιο ότι οι μέθοδοι μας και γενικότερα τα αποτελέσματα που εξήγαμε επεκτείνονται και σε άλλες κατανομές θορύβων. Το σημαντικό σημείο που πρέπει να προσέξουμε εδώ είναι να περιορίσουμε τα υπόλοιπα έτσι ώστε το πραγματικό διάνυσμα β να είναι εφικτό για το πρόβλημα βελτιστοποίησης. Συγκεκριμένα, αυτό σημαίνει ότι θα οριστεί το λ_p ώστε το $Z^* = sup_i|\langle X^i, z\rangle|$ να είναι μικρότερο του $\lambda_p \cdot \sigma$ με μεγάλη πιθανότητα. Έχουμε $z \sim N(0, \sigma^2 I_n)$ το οποίο επιτυγχάνεται για $\lambda_p = \sqrt{2logp}\ \cdot$ αν και θα μπορούσαμε να επιτύχουμε παρόμοια αποτελέσματα και με διαφορετικούς τύπους.

5.7 Εφαρμογή στην επεξεργασία σήματος

Σε αυτή την παράγραφο θα ασχοληθούμε με την ανάκτηση ενός μονοδιάστατου σήματος (*signal*) $f \in R^p$ (Candès και Tao (2007)) από θορυβώδης και υποδειγματοληφθέντες (*undersampled*) συντελεστές Fourier της μορφής:

$$y_j =< f, \varphi_j > +z_j \qquad , 1 \le j \le n \tag{5.32}$$

όπου $\varphi_j(t),\ t = 0, \dots, p-1$, είναι ημιτονοειδής κύμα της μορφής

$$\varphi_j(t) = \sqrt{2/n}\cos(\pi(k_j + 1/2)(t + 1/2)) \tag{5.33}$$

, $k_j \in \{0,1, \dots, p-1\}$.

Θεωρούμε το σήμα f του Σχήματος 5.1 το οποίο δεν είναι αραιό, αλλά η ακολουθία των συντελεστών των κυματιδίων β είναι. Κατά συνέπεια μπορούμε εξίσου καλά να εκτιμήσουμε τους συντελεστές του σε μια ωραία βάση κυματιδίων.

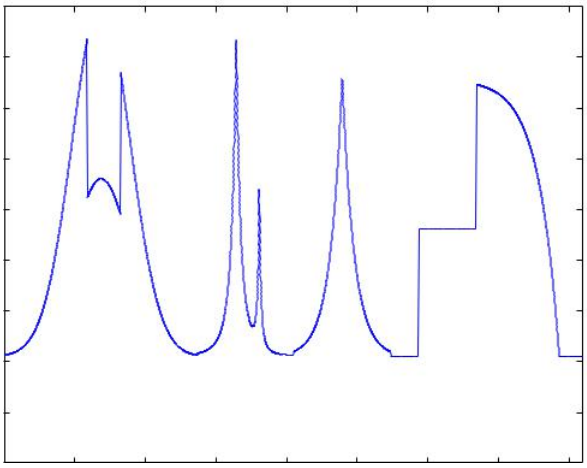
Ορίζουμε το Φ να είναι ο πίνακας όπου φ_k να είναι οι γραμμές του και τον W να είναι ο ορθογώνιος πίνακας κυματιδίων όπου οι στήλες του να αποτελούνται από τα κυματίδια. Τότε θα έχουμε

$$Y = Xb + z \text{ , } \textit{όπου } X = \Phi W \tag{5.34}$$

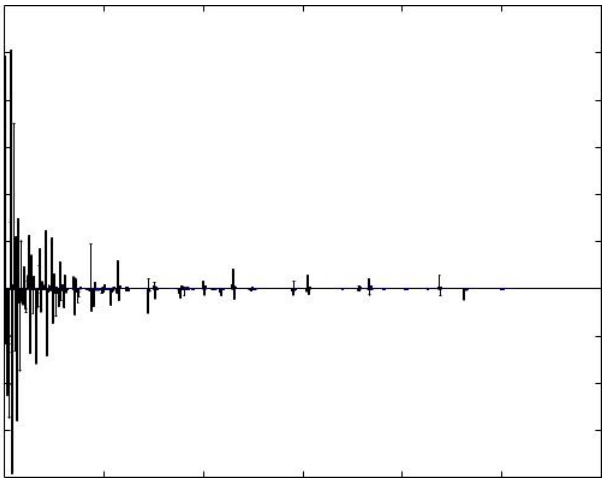
και η διαδικασία εκτίμησης μας εφαρμόζεται ως έχει.

Πίνακας 5.1: Παρουσιάζεται η απόδοση της διαδικασίας Gauss-Dantzig στην εκτίμηση ενός σήματος από υποδειγματοληφθέντες και θορυβώδεις συντελεστές Fourier. Το υποσύνολο των μεταβλητών εδώ εκτιμάται μέσω $|\hat{\beta}_i| > \sigma/4$, με το $\hat{\beta}$ να είναι όπως αυτό της σχέσης (5.7).

SNR $a = \frac{\|X\beta\|}{\sqrt{n\sigma^2}}$	100	20	10	2	1	0.5
$\sum_i (\hat{\beta}_i - \beta_i)^2 / \sum_i \min(\beta_i^2, \sigma^2)$	15.51	2.08	1.40	1.47	0.91	1.00



Σχήμα 5.1: Αναπαρίσταται το σήμα f μίας διάστασης.



Σχήμα 5.2: Οι πρώτοι 512 συντελεστές κυματιδίων του f.

Το σήμα δοκιμής έχει μέγεθος $p = 4{,}096$ (Σχήμα 5.1 , 5.2), και θα επιλέξουμε ένα σύνολο συχνοτήτων μεγέθους $n = 512$ εξάγοντας τις 128 χαμηλότερες συχνότητες και επιλέγοντας τυχαία τις υπόλοιπες. Με αυτό το σύνολο των παρατηρήσεων, ο στόχος είναι να μελετηθεί η ποσοτική συμπεριφορά της διαδικασίας του Gauss-Dantzig selector για διάφορα επίπεδα θορύβου. Εξαιτίας του παράγοντα $\sqrt{2/n}$ στο ορισμό της $\varphi_j(t)$, οι στήλες του πίνακα X έχουν μέγεθος περίπου 1 και για κάθε στήλη, τα μεμονωμένα κατώτατα όρια λ_i, $|(X^*r)_i| \leq \lambda_i \cdot \sigma$, καθορίζονται με την εξέταση της εμπειρικής κατανομής του $|(X^*z)_i|$. Έχουμε προσαρμόσει το σ έτσι ώστε $\alpha^2 = \frac{\|X\beta\|_{l_2}^2}{E\|z\|_{l_2}^2} = \frac{\|X\beta\|_{l_2}^2}{n\sigma^2}$ για διάφορα επίπεδα του λόγου α του σήματος προς τον θόρυβο. Χρησιμοποιούμε τα κυματίδια Daubechies[18] με 4 μηδενιζόμενες ροπές (*vanishing moments*)[19] για την αναπαράσταση. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.1. Όπως μπορεί

[18] Οι συναρτήσεις κυματιδίων Daubechies έχουν πεπερασμένο διάστημα ορισμού $[-K, K]$, που με τη σειρά του οδηγεί σε συζυγή φίλτρα με πεπερασμένο αριθμό συντελεστών. Η συνάρτηση κλιμάκωσης δεν ορίζεται αναλυτικά αλλά υπολογίζεται με βάση μια αναδρομική σχέση. Είναι επίσης βέλτιστα με την έννοια ότι έχουν το μέγιστο αριθμό μηδενικών ροπών για δεδομένο ενεργό πεδίο ορισμού. Η Ingrid Daubechies απέδειξε ότι αν μια συνάρτηση έχει p μηδενικές ροπές τότε υποχρεωτικά έχει μήκος $2p - 1$.

[19] Οι μηδενιζόμενες ροπές είναι σημαντικές για την μέτρηση της τοπικής κανονικότητας (local regularity) ενός σήματος. Δηλαδή, η τάξη μετασχηματισμού ενός κυματιδίου δίνεται από τον αριθμό των μηδενιζόμενων ροπών της ανάλυσης του κυματιδίου.

κανείς να δει, η υψηλή στατιστική ακρίβεια ισχύει σε ένα ευρύ φάσμα των επιπέδων του θορύβου. Παρουσιάζει ενδιαφέρον το γεγονός ότι ο εκτιμητής είναι λιγότερο ακριβής όταν το επίπεδο θορύβου είναι πολύ μικρό ($\alpha = 100$) , το οποίο δεν αποτελεί έκπληξη δεδομένου ότι σε αυτή την περίπτωση υπάρχουν 178 συντελεστές κυματιδίων που υπερβαίνουν το σ σε απόλυτη τιμή .

Στο τελευταίο μας παράδειγμα, εξετάζουμε το πρόβλημα της αναπαράστασης μιας εικόνας από υποδειγματοληφθέντες συντελεστές Fourier. Σε αυτή την περίπτωση το $\beta(t_1, t_2),\ 0 \le t_1,\ t_2 < N,$ είναι μια άγνωστη N επί N εικόνα και έτσι αυτό το p είναι ο αριθμός των αγνώστων pixels, $p = N^2$. Ως συνήθως, τα δεδομένα δίνονται από την σχέση $y = Xb + z$, όπου

$$(X\beta)_k = \sum_{t_1,t_2} \beta(t_1, t_2)\cos(2\pi(k_1 t_1 + k_2 t_2)/N), \quad k = (k_1, k_2)$$
$$\text{ή} \quad (X\beta)_k = \sum_{t_1,t_2} \beta(t_1, t_2)\sin(2\pi(k_1 t_1 + k_2 t_2)/N). \tag{5.35}$$

Στο παράδειγμα μας, η εικόνα β δεν είναι αραιή, αλλά η κλίση είναι . Επομένως, για να αναπαρασταθεί η εικόνα, εφαρμόζουμε της στρατιγική εκτίμισης μας και ελαχιστοποιούμε την l_1-νόρμα του μεγέθους της κλίσης γνωστό και ως συνολική μεταβολή (*total variation*) του β,

$$min\left\|\tilde{\beta}\right\|_{TV} \text{ δεδομένου ότι } |(X^* r)_i| \le \lambda_i \cdot \sigma \tag{5.36}$$

Η συνολική μεταβολή της νόρμας θα είναι της μορφής

$$\left\|\tilde{\beta}\right\|_{TV} = \sum_{t_1,t_2} \sqrt{|D_1\tilde{\beta}(t_1, t_2)|^2 + |D_2\tilde{\beta}(t_1, t_2)|^2} \tag{5.37}$$

όπου το D_1 είναι η πεπερασμένη διαφορά $D_1\tilde{\beta} = \beta(t_1, t_2) - \beta(t_1 - 1, t_2)$ και $D_2\tilde{\beta} = \beta(t_1, t_2) - \beta(t_1, t_2 - 1)$. Με λίγα λόγια $\left\|\tilde{\beta}\right\|_{TV}$ είναι η l_1-νόρμα του μεγέθους της κλίσης $D\tilde{\beta} = \left(D_1, \tilde{\beta}\right) - \left(D_2, \tilde{\beta}\right)$.

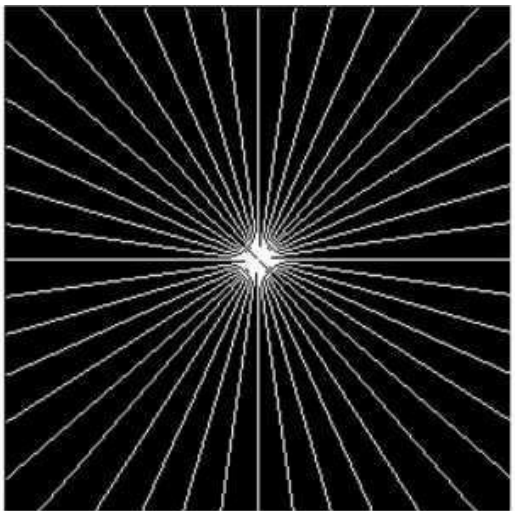
Το παράδειγμα μας ακολουθεί τα πρότυπα απόκτησης δεδομένων πολλών συσκευών πραγματικής απεικόνισης (*real imaging devices*) που μπορούν να συλλέξουν δείγματα υψηλής ευκρίνειας κατά μήκος ακτινικών γραμμών σε σχετικά λίγες γωνίες. Το Σχήμα 5.3 απεικονίζει μια χαρακτηριστική περίπτωση όπου ένα συγκεντρώνει $N = 256$ δείγματα κατά μήκος καθεμιάς από τις 22 ακτινικές γραμμές (*radial lines*).

Στη συνέχεια σε ένα πρώτο πείραμα, παρατηρούμε $22 * 256$ πραγματικές τιμές θορυβώδων συντελεστών Fourier και τη χρήση της σχέσης (5.36) για την αναπαράσταση του προβλήματος που απεικονίζεται στα Σχήματα 5.3-5.6 . Ο αριθμός των παρατηρήσεων τότε είναι $n = 156$, ενώ υπάρχουν $p = 65{,}536$ παρατηρήσεις. Με άλλα λόγια περίπου το $91{,}5\%$ των δισδιάστατων (2D) συντελεστών Fourier του β λείπουν.

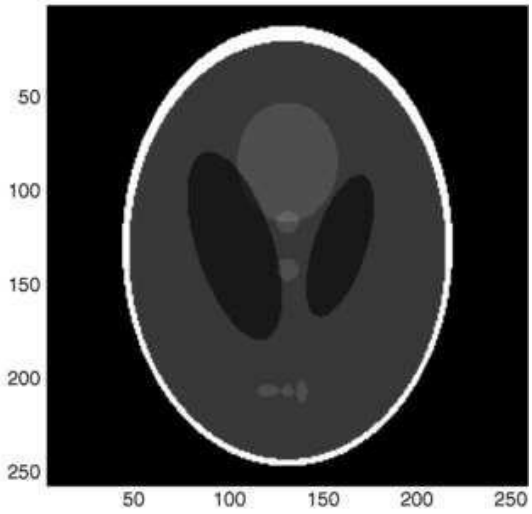
Το SNR (*Signal to Noise Ratio*) [20] σε αυτό το πείραμα είναι ισοδύναμο με $\frac{\|X\beta\|_{l_2}}{\|z\|_{l_2}} = 5.85$. Το Σχήμα 5.5 δείχνει την αναπαράσταση που λαμβάνεται με τον καθορισμό των απαρατήρητων συντελεστών Fourier στο μηδέν, ενώ το Σχήμα 5.6 δείχνει την αναπαράσταση (5.36).

Συνεχίζουμε με ένα δεύτερο παράδειγμα όπου η άγνωστη εικόνα είναι τώρα 512 επί 512 έτσι ώστε το $p = 262{,}144$ και το $n = 22 * 512 = 11{,}264$. Το κλάσμα των συντελεστών Fourier που λείπουν, πλησιάζει τώρα το 96%. Ο λόγος SNR είναι περίπου ο ίδιος με πριν, $\frac{\|X\beta\|_{l_2}}{\|z\|_{l_2}} = 5.77$. Οι αναπαραστάσεις είναι πολύ καλής ποιότητας, ειδικά σε σύγκριση με την απλοϊκή αναπαράσταση η οποία ελαχιστοποιεί την ενέργεια της αναπαράστασης που υπόκεινται σε αντιστοιχία τα παρατηρούμενα δεδομένα.

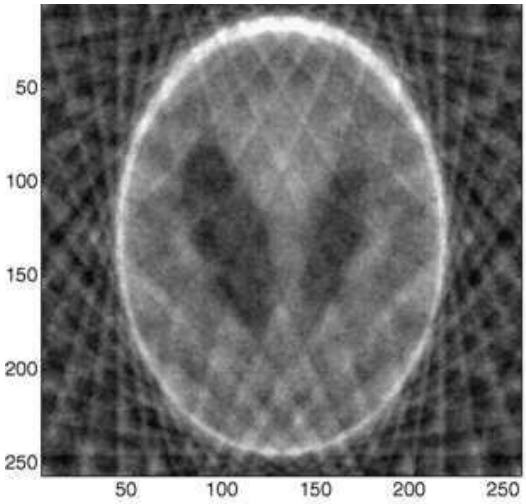
[20] Εκφράζει το λόγο της ισχύος του σήματος προς την ισχύ του θορύβου και δίδεται από την εξίσωση $SNR = 10 * log10\,(S / N)$ όπου προφανώς S είναι η ισχύς του σήματος και N η ισχύς του θορύβου. Ο βασικός ρόλος του *SNR* είναι η δυνατότητα που μας παρέχει να μετρήσουμε το μέγιστο ρυθμό μεταφοράς δεδομένων του καναλιού, υπό συνθήκες παρουσίας θορύβου.



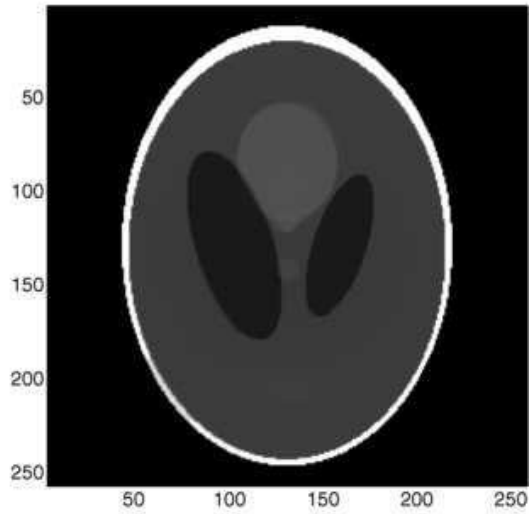
Σχήμα 5.3 : Δειγματοληψία διαίρεσης στο επίπεδο συχνότητας. Οι συντελεστές Fourier αποτελούν δείγματα κατά μήκος 22 περίπου ακτινικών γραμμών και σε αυτή την περίπτωση $n \approx 0.086p$.



Σχήμα 5.4: Το ψηφιακό ομοίωμα του Logan-Sheep.



Σχήμα 5.5 : Αναπαράσταση της ελάχιστης ενέργειας που προκύπτει θέτοντας με μηδέν τους απαρατήρητους (*unobserved*) συντελεστές Fourier.



Σχήμα 5.6: Αναπαράσταση που προκύπτει με την ελαχιστοποίηση της ολικής κύμανσης (*total-variation*) που δίνεται από την σχέση (5.36)

ΚΕΦΑΛΑΙΟ 6

DANTZIG SELECTOR
ΣΕ ΜΕΡΙΚΩΣ ΓΡΑΜΜΙΚΑ ΜΟΝΤΕΛΑ

6.1 Εισαγωγή

Σε αυτή την ενότητα, θα ασχοληθούμε με την επιλογή μεταβλητής και την παραμετρική εκτίμηση των μερικώς γραμμικών μοντέλων (*partially linear models*) διαμέσου του Dantzig selector. Όσων αφορά τις ασυμπτωτικές ιδιότητες ενός μεγάλου δείγματος, μελετήθηκαν όταν το μέγεθος του δείγματος n τείνει στο άπειρο καθώς το p είναι σταθερό. Όμως θα δούμε ότι πιθανώς ο Dantzig selector να μην είναι συνεπής · για αυτό και θα αναφερθούμε στον adaptive (προσαρμοστικό) Dantzig selector των Dicker και Li (2009).

Επιπλέον, λαμβάνουμε υπόψη το γεγονός ότι ο adaptive Dantzig selector για παραμετρική συνιστώσα των μερικώς παραμετρικών μοντέλων υιοθετεί, υπό κάποιες κατάλληλες προϋποθέσεις, τις ιδιότητες oracle. Τόσο ο adaptive Dantzig selector όσο και ο Dantzig selector βελτιστοποίησης μπορούν να εφαρμοστούν μέσω του αλγόριθμου DASSO που προτάθηκε από τους James, Radchenkob και Lv (2009).

Όσον αφορά τα μερικώς γραμμικά μοντέλα που θα ασχοληθούμε, γνωρίζουμε ότι αποτελούν μια ευρεία κατηγορία ημιπαραμετρικών μοντέλων που είναι καλά ερμηνεύσιμη δεδομένου ότι περιέχει τόσο παραμετρικά όσο και μη παραμετρικά στοιχεία. Τα μοντέλα αυτά επιτρέπουν την ευκολότερη ερμηνεία της επίδρασης κάθε μεταβλητής και μπορούν να προτιμηθούν από μια εντελώς μη παραμετρική παλινδρόμηση λόγω της γνωστής «κατάρας της διαστατικότητας»

(*curse of dimensionality*) [21]. Η επιλογή του μοντέλου για τα μερικώς γραμμικά μοντέλα έχει μια περεταίρω δυσκολία μιας και πρέπει να εκτιμήσουμε την μη παραμετρική συνάρτηση καθώς και τα άλλα παραμετρικά προβλήματα· όπως είναι η επιλογή του εύρους ζώνης (*bandwidth*) αλλά και η επιλογή της παραμέτρου tuning, λ.

6.2 Dantzig selector για μερικώς γραμμικά μοντέλα

Υποθέτουμε ότι το σύνολο $\{(Y_i, X_i, T_i)\,, i = 1,2, \dots, n\}$ είναι ένα τυχαίο δείγμα του μερικώς γραμμικού μοντέλου

$$Y = X'\beta + g(T) + \varepsilon \qquad (6.1)$$

όπου X είναι ένα διάνυσμα p-διαστάσεων ανεξάρτητων μεταβλητών, το β είναι αραιό διάνυσμα γεγονός που σημαίνει ότι μόνο κάποιες από τις συνιστώσες του είναι μη μηδενικές, $g(\cdot)$ είναι μια άγνωστη ομαλή συνάρτηση της βοηθητικής μονομεταβλητής T. Σε αυτή την παράγραφο θα ασχοληθούμε κυρίως με την μονομεταβλητή T. Η μέθοδος που θα προταθεί ισχύει και για την πολυμεταβλητή T, εντούτοις η επέκταση στη περίπτωση των πολυμεταβλητών πιθανό να είναι πρακτικά λιγότερο χρήσιμη λόγω της «κατάρας της διαστατικότητας».

Υποθέτουμε ότι η T είναι τυχαία και παίρνει τιμές σε ένα συμπαγές διάστημα όπου για λόγους απλότητας θα θεωρήσουμε ότι αυτό θα είναι το [0,1]. Επιπλέον να σημειωθεί ότι Y είναι η μεταβλητή απόκρισης και ε είναι το ανεξάρτητο τυχαίο σφάλμα του (X, T) με μέση τιμή να είναι 0 και σταθερή απόκλιση σ.

Έστω $X = (X_1, X_2, \dots, X_n)'$ να είναι ο πίνακας σχεδιασμού, όπου $X_i = (X_{i1}, X_{i2}, \dots, X_{in})'\,, i = 1,2, \dots, n.$ Έστω $T = (T_1, T_2, \dots, T_n)'$, $Y = (Y_1, Y_2, \dots, Y_n)'$, $g(T) = (g(T_1), g(T_2), \dots, g(T_n))'$, $\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$. Ακολούθως, υποθέτουμε ότι X_{ij} και T_i συσχετίζονται μέσω του μοντέλου παλινδρόμησης $E(X_{ij}|T_i) =$

[21] Ο όρος αυτός εισήχθη το 1961 από τον Bellman. Αναφέρεται στο πρόβλημα της ανάλυσης δεδομένων πολλών μεταβλητών καθώς αυξάνει η διάσταση. Στην πραγματικότητα η κατάρα της διαστατικότητας σημαίνει ότι για δεδομένο αριθμό δειγμάτων, υπάρχει μια μεγάλη διάσταση των χαρακτηριστικών διανυσμάτων πάνω από την οποία η απόδοση του ταξινομητή μας θα μειώνεται.

$q_j(T_i)$, όπου $q_j(T_i)$ $(1 \le j \le p)$ είναι ομαλές συναρτήσεις με δυο συνεχείς παράγωγους και το T_i $(1 \le i \le n)$ παράγεται από το T το οποίο έχει συνάρτηση πυκνότητας[22] $p(t)$. Υποθέτουμε επίσης ότι $X_{ij} = q_j(T_i) + n_{ij}$ $(1 \le i \le n, 1 \le j \le p)$, όπου τα n_{ij} ικανοποιούν την $E(n_{ij}|T_i) = 0$ και είναι ανεξάρτητα από τα ε_i. Τότε ο πίνακας X μπορεί να αναλυθεί ως $X = q + n$, όπου $q = (q_{ij})_{n\times p}$, $q_{ij} = q_j(T_i)$, $\eta = (\eta_{ij})_{n\times p}$.

Υποθέτουμε ότι το διάνυσμα του συντελεστή, β ,είναι υψηλών διαστάσεων και αραιό. Έστω ότι $C^* = \{j, \beta_j \neq 0\}$, όπου β_j ορίζεται ως η j-οστή συνιστώσα του β. Αναφερόμαστε στο C^* ως το πραγματικό μοντέλο. Ο κύριος στόχος της επιλογής του μοντέλου και της εκτίμησης του είναι ο προσδιορισμός του C^* και η εκτίμηση των μη μηδενικών συνιστωσών του β.

Από τη σχέση (6.1) φαίνεται ότι $E(Y|T) = E(X|T)'\beta + g(T)$ και άρα θα έχουμε ότι

$$E([X - E(X|T)][\big(Y - E(Y|T)\big) - (X - E(X|T))'\beta]) = 0 \qquad (6.2)$$

Ορίζουμε ότι $m_X(T) = E(X|T)$, $m_Y(T) = E(Y|T)$. Έστω ότι οι $\widehat{m}_X(T)$, $\widehat{m}_Y(T)$ είναι οι εκτιμητές των $m_X(T)$ και $m_Y(T)$ αντίστοιχα. Έπειτα ορίζουμε τον Dantzig selector ως:

$min_{\beta\in R^p}(\|\beta\|_1)$ έτσι ώστε να ισχύει

$$max_{1\le j\le p}\left|\left(\sum_{i=1}^{n}(X_i - \widehat{m}_X(T_i))(Y_i - \widehat{m}_Y(T_i) - (X_i - \widehat{m}_X(T_i))'\beta)\right)_j\right| \le \lambda \cdot \sigma \qquad (6.3)$$

όπου λ είναι η παράμετρος *tuning*.

[22] Γενικά, η συνάρτηση πυκνότητας για συνεχείς τυχαίες μεταβλητές ορίζεται από τη σχέση $f(x)dx = P(x < X \le X + dx)$. Δηλαδή μας δίνει την πιθανότητα η μεταβλητή X να πάρει τιμές σε ένα ορισμένο απειροστό διάστημα γύρω από το σημείο x.

Η εκτίμηση των $m_X(T)$ και $m_Y(T)$ γίνεται με τη χρήση της εκτίμησης N-W kernel [23]. Συγκεκριμένα προκύπτει ότι οι αντίστοιχοι εκτιμητές τους θα είναι:

$$\widehat{m}_X(T) = \frac{\sum_{i=1}^{n} K\left(\frac{T_i - T}{h}\right) X_i}{\sum_{i=1}^{n} K\left(\frac{T_i - T}{h}\right)} \quad \text{και} \quad \widehat{m}_Y(T) = \frac{\sum_{i=1}^{n} K\left(\frac{T_i - T}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{T_i - T}{h}\right)} \tag{6.4}$$

όπου $K(\cdot)$ είναι η συνάρτηση *kernel* και h είναι το εύρος ζώνης.

Η βέλτιστη τιμή του β για την λύση του προβλήματος (6.3) είναι απλά η εκτίμηση του β που συμβολίζεται ως $\hat{\beta}$. Συνήθως, οι συντελεστές θα εκτιμώνται με μηδέν όταν είναι ίσοι με μηδέν ή όταν τείνουν στο μηδέν. Συνεπώς, μπορούμε να πάρουμε την εκτίμηση του μη παραμετρικού στοιχείου $g(T)$ ως

$$\hat{g}(T) = \widehat{m}_Y(T) - (\widehat{m}_X(T))'\hat{\beta} \tag{6.5}$$

Ορίζουμε τα ακόλουθα μεγέθη ως $\tilde{Y}_i = Y_i - m_Y(T_i)$, $\tilde{X}_i = X_i - m_X(T_i)$, $\hat{Y}_i = Y_i - \widehat{m}_Y(T_i)$, $\hat{X}_i = X_i - \widehat{m}_X(T_i), \tilde{X} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)'$, $\tilde{Y} = (\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_n)'$, $\hat{X} = (\hat{X}_1, \hat{X}_2, \dots, \hat{X}_n)'$, $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n)'$, $\hat{g} = (1 - K)g(T)$, $\hat{q} = (1 - K)q$, $\hat{\eta} = (1 - K)\eta$ όπου $K = (K_{ij})_{n \times n}$, $K_{ij} = \frac{K\left(\frac{T_i - T_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{T_i - T_j}{h}\right)}$.

Τότε η σχέση (6.3) μπορεί να αναδιατυπωθεί ως:

$$min_{\beta \in R^p (\|\beta\|_1)} \text{ έτσι ώστε } \left\|\hat{X}'\left(\hat{Y} - \hat{X}\beta\right)\right\|_\infty \leq \lambda \cdot \sigma \tag{6.6}$$

Αυτό είναι ένα γραμμικό πρόγραμμα και έτσι πλέον μπορεί εύκολα να χρησιμοποιηθεί το σύνηθες λογισμικό. Παρόλα αυτά ο αλγόριθμος μπορεί να γίνει υπο-

[23] Η παλινδρόμηση Kernel είναι μια μη παραμετρική τεχνική στη στατιστική για εκτίμηση. Το 1964 οι Nadaraya και Watson πρότειναν τους εκτιμητές m χρησιμοποιώντας την kernel ως μια συνάρτηση στάθμισης.

λογιστικά δαπανηρός εάν το σύνολο των δεδομένων είναι μεγάλο. Ο αλγόριθμος DASSO, που σχετίζεται με τον LARS (*Least-Angle Regression*) [24], μπορεί να βρει αποτελεσματικά τις εκτιμήσεις του Dantzig selector για όλες τις τιμές της παραμέτρου tuning λ και επίσης έχει παρόμοιο υπολογιστικό κόστος με τον LARS. Χρησιμοποιώντας $\hat{Y}$ και $\hat{X}$ αντί Y και X αντίστοιχα, ο αλγόριθμος DASSO μπορεί επίσης να χρησιμοποιηθεί για να βρει τις εκτιμήσεις των συντελεστών στα μερικώς γραμμικά μοντέλα για όλες τις τιμές του λ αφότου κάνει μια μικρή αλλαγή στις μεταβλητές.

Ένα σημαντικό ζήτημα που παρουσιάζεται είναι η επιλογή του h και λ. Σε αντίθεση με την επιλογή μεταβλητής για γραμμικά μοντέλα, εδώ χρειαζόμαστε να επιλέξουμε δύο παραμέτρους, την h και λ. Για να βρεθούν οι βέλτιστες τιμές αυτών των παραμέτρων, μια κοινή προσέγγιση θα μπορούσε να ήταν η αναζήτηση ενός δισδιάστατου (*2D*) πλέγματος, χρησιμοποιώντας μερικά κριτήρια data-driven, όπως είναι το CV ή το GCV (Craven and Wahba (1979)), τα οποία πιθανό να είναι υπολογιστικά χρονοβόρα. Κατά συνέπεια, εδώ μπορούμε να επιλέξουμε το h και το λ όπως έκαναν οι Wang, Li και Tsai (2007). Συγκεκριμένα, πρώτα από όλα προσαρμόζουμε την προσέγγιση των Fan και Li (2004) για να λάβουμε τον εκτιμητή $\hat{\beta}_{DF}$ του β για το πλήρες μοντέλο, ο οποίος είναι συνεπής εκτιμητής ως προς το $\sqrt{n}$,που προτάθηκε από τους Yatchew (1997) και Wang, Brown και Cai (2011). Έπειτα, αφότου αντικατασταθεί το β από το $\hat{\beta}_{DF}$ στο μοντέλο (6.1), τότε το μοντέλο που προκύπτει θα είναι ένα ομαλό πρόβλημα μιας διάστασης (1D) και επομένως η επιλογή του εύρους ζώνης μπορεί να γίνει ως συνήθως. Για την επιλογή του εύρους ζώνης μπορούμε να χρησιμοποιήσουμε το CV , το GVC ή την μέθοδο plug-in από τους Ruppert, Sheather και Wand (1995). Τέλος, από την επιλογή του εύρους ζώνης h θα υπάρχει μόνο μια παράμετρος tuning λ η οποία επιλέγεται. Επομένως, μπορούν να χρησιμοποιηθούν τα CV, GCV ,AIC και BIC.

[24] Είναι ένας αλγόριθμος που στη γενική του μορφή χρησιμοποιείται για την αναζήτηση εκτιμητών σε συστήματα γραμμικών εξισώσεων όπου ο αριθμός τους είναι μεγαλύτερος από το πλήθος των άγνωστων και κατά συνέπεια δεν υπάρχει ακριβής λύση. Προτάθηκε από τους Bradley Efron, Trevor Hastie, Iain Johnstone και Robert Tibshirani.

6.3 Θεωρητικά αποτελέσματα του Dantzig selector

Προτού εξετάσουμε τις ασυμπτωτικές ιδιότητες του Dantzig selector, θα κάνουμε μια συνοπτική αναφορά στα υποδιανύσματα (*sub-vectors*) των διανυσμάτων και στους υποπίνακες (*sub-matrices*) των πινάκων. Δεδομένου του υποσυνόλου $C \subseteq \{1, \dots, p\}$ και $\beta \in R^p$, ορίζουμε $|C|$ την πληθικότητα του C και $\beta_C = (\beta_j)_{j \in C}$ να είναι το διάνυσμα $|C| \times 1$ του οποίου τα στοιχεία είναι εκείνα του β με δείχτες του C. Ομοίως, ορίζουμε τον υποπίνακα X_C να είναι ο $n \times |C|$ ρπίνακας του οποίου οι στήλες είναι αυτές του X με δείκτες από το C. Για τα υποσύνολα $C_1, C_2 \subseteq \{1, \dots, p\}$ και το $p \times p$ πίνακα Σ, ορίζουμε Σ_{C_1,C_2} να είναι ο $|C_1| \times |C_2|$ πίνακας που δημιουργήθηκε από τις γραμμές που εξήχθησαν σύμφωνα με το C_1 και τις στήλες σύμφωνα με το C_2. Ας ορίσουμε το $\bar{C}$ να είναι το συμπλήρωμα του C.

Ακολούθως, να τονισθεί ότι σε όλα τα ασυμπτωτικά αποτελέσματα υποθέτουμε ότι το p είναι σταθερό και το n τείνει στο άπειρο. Επίσης, ορίζεται ο πίνακας Σ ως εξής $\Sigma = E\big(X - E(X|T)\big)\big(X - E(X|T)\big)' = (\sigma_{ij})_{p \times p}$. Διαιρώντας τη σχέση (6.6) με n και δεδομένου ότι $n \to \infty$ θα έχουμε:

$$\frac{1}{n}\left\|\hat{X}'\left(\hat{Y} - \hat{X}\hat{\beta}\right)\right\|_\infty = \left\|\Sigma\left(\beta - \hat{\beta}\right) + o_p(1)\right\|_\infty \tag{6.7}$$

Επομένως, ο περιορισμός μπορεί να αναδιατυπωθεί ως εξής:

$$\left\|\Sigma\left(\beta - \hat{\beta}\right) + o_p(1)\right\|_\infty \leq \frac{\lambda \cdot \sigma}{n} \tag{6.8}$$

Ακολουθούν δύο προτάσεις που διατυπώθηκαν από τους Dicker και Lin (2009) και θα μας χρησιμεύσουν για την συνέχεια. Η Πρόταση 6.1 αναφέρεται στις ιδιότητες μοναδικότητας του Dantzig selector. Η Πρόταση 6.2 διατυπώνει ότι τόσο ο Dantzig selector όσο και η LASSO έχουν μοναδική λύση στη συντριπτική πλειοψηφία των περιπτώσεων.

Πρόταση 6.1

(α) Έστω ότι υπάρχουν τα υποσύνολα $A, B \subseteq \{1, \dots, p\}$, και τα διανύσματα $\mu^0 \in R^{|B|}$, $\beta^0 \in F$. Εάν ισχύουν οι πιο κάτω ιδιότητες , τότε ο Dantzig selector έχει πολλαπλές λύσεις.

1. $\|n^{-1}X^T X_B \mu^0\|_\infty \leq 1, n^{-1}X_A^T X_B \mu^0 \in \{\pm 1\}^{|A|}$ και $\dim[null(n^{-1}X_B^T X_A)] > 0$.

2. $n^{-1}\beta^T X^T X_B \mu^0 \geq \|\beta^0\|_1$ για όλα τα $\beta \in F$.

3. $A = \{j : \ \beta_j^0 \neq 0\}$.

4. $n^{-1}\left|X_j^T(y - X\beta^0)\right| = \lambda$ για όλα τα $j \ \in B$ και $n^{-1}\left|X_j^T(y - X\beta^0)\right| < \lambda$ για όλα τα $j \notin B$.

(β) Εάν το εφικτό σύνολο του Dantzig selector, F, δεν είναι παράλληλο με την l^1-μπάλα, τότε ο Dantzig selector έχει μοναδική λύση.

(γ) Υποθέτουμε ότι το $\lambda > 0$ και ότι η LASSO έχει πολλαπλές λύσεις. Τότε υπάρχει ένα υποσύνολο $A \subseteq \{1, \dots, p\}$ και το διάνυσμα $w \in R^p$ έτσι ώστε $\|n^{-1}X^T Xw\|_\infty \leq 1$, $n^{-1}X_A^T Xw \ \in \{\pm 1\}^{|A|}$ και $dim[null(n^{-1}X^T X_A)] > 0$.

■

Πρόταση 6.2

Υποθέτουμε ότι τα $x_1, \dots, x_n$ είναι ανεξάρτητα και προέρχονται από μια συνεχή κατανομή που υπακούνε στο μέτρο Lebesgue στον R^p. Τότε ο $n^{-1}X^T X$ είναι παράλληλος με την l^1-μπάλα με πιθανότητα 0. Κατά συνέπεια ο Dantzig selector και η LASSO έχουν μοναδική λύση με πιθανότητα 1.

■

Παρόμοια με τις Προτάσεις 6.1 και 6.2 μπορούμε εύκολα να λάβουμε ότι ο Dantzig selector (6.6) έχει την μοναδική λύση όταν ο πίνακας Σ δεν είναι παράλληλος με την l_1-μπάλα, η οποία ορίζεται ως $B_1 = \{u \in R^p : \|u\|_1 = 1\}$,και όταν ισχύουν οι ακόλουθες συνθήκες κανονικότητας :

(α) Η kernel συνάρτηση $K(\cdot)$ είναι συμμετρική γύρω από το μηδέν με το φορέα (*support*) στο διάστημα.

(β) Το εύρος ζώνης h στην εκτίμηση του $m_X(T)$ και του $m_Y(T)$ παίρνει την τάξη του $O(n^{-1/5})$.

(γ) Η συνάρτηση πυκνότητας $p(T)$ του T είναι φραγμένη μακριά από το μηδέν και έχει φραγμένη συνεχή δεύτερη παράγωγο.

(δ) Οι $m_X(T)$ και $m_Y(T)$ έχουν φραγμένες και συνεχείς δεύτερες παραγώγους .

(ε) Το ίχνος (*trace*) $(K'K) = \sum_{i=1}^{n}\sum_{j=1}^{n} K_{ij}^2 = O_p\left(n^{\frac{1}{5}}\right)$.

(στ) $\hat{q}_j(T_i) = O_p(n^{-2/5})$ για κάθε $i = 1, \dots, n$ και $j = 1, \dots, p$.

(η) $\|(I-K)g(T)\|^2 = \|\hat{g}\|^2 = O_p(n^{1/5})$.

(θ) Ο πίνακας Σ είναι θετικά ορισμένος.

■

Τώρα, σχετικά με τον παραλληλισμό που αναλύθηκε πιο πάνω διαδραματίζει ένα βασικό ρόλο στην έρευνα των ασυμπτωτικών ιδιοτήτων του εκτιμητή Dantzig selector. Ο καθορισμός ενός πίνακα που είναι παράλληλος με την l_1-μπάλα δίνεται από την ακόλουθη γενική πρόταση που εξήγαγαν οι Dicker και Lin (2009):

(α) Έστω C είναι ένας $p \times p$ συμμετρικός πίνακας. Ο πίνακας C είναι παράλληλος με την l_1-μπάλα εάν και μόνο ισχύει η ακόλουθη συνθήκη. Υπάρχουν τα υποσύνολα $A, B \subseteq \{1, \dots, p\}$ και το διάνυσμα $\boldsymbol{w} \in R^{|B|}$ έτσι ώστε $\|C_B \boldsymbol{w}\|_\infty \leq 1$, $C_{A,B} w \in \{\pm 1\}^{|A|}$, και dim$[null(C_{A,B})] > 0$.

(β) Το εφικτό σύνολο για τον Dantzig selector, F, είναι παράλληλο με την l_1-μπάλα εάν και μόνο η ποσότητα $n^{-1}X^TX$ είναι παράλληλη με την l_1-μπάλα.

■

Ακολουθεί ένα παράδειγμα που περιγράφει τον παραλληλισμό που μόλις αναλύθηκε . Έστω ότι εξετάζουμε την περίπτωση το $p = 2$ για το μοντέλο $Y = X'\beta + g(T) + \varepsilon$. Το εφικτό σύνολο του Dantzig selector (6.6) είναι $D = \left\{ \beta : \left\| \hat{X}'(\hat{Y} - \hat{X}\hat{\beta}) \right\|_\infty \leq \lambda \cdot \sigma \right\}$, και οι λύσεις στον Dantzig selector είναι τα σημεία $\beta \in D$ της ελάχιστης l_1-νόρμας. Έστω ότι $B_\gamma = \{u \in R^2 : \|u\|_1 \leq \gamma\}$. Δεδομένου του λ, η λύση στον Dantzig selector είναι τα πρώτα σημεία της τομής μεταξύ του D και του B_γ, δηλαδή $D \cap B_\gamma$, καθώς το γ αυξάνεται και $\gamma \geq 0$. Επίσης, γνωρίζουμε ότι η μονοδιάστατη (*1D*) επιφάνεια του B_γ έχει κλίση ± 1 , και ο Dantzig selector θα έχει πολλαπλές λύσεις μόνο όταν η μονοδιάστατη (*1D*) επιφάνεια του D έχει επίσης κλίση ± 1. Με λίγα λόγια να ισχύει ότι το D θα είναι παράλληλο με την l_1-μπάλα.

Στη συνέχεια, υποθέτοντας ότι το Σ δεν είναι παράλληλο με την l^1-μπάλα, εάν $\frac{\lambda\sigma}{n} \to 0$ θα δούμε ότι $\hat{\beta} \xrightarrow{P} \beta_0$. Εάν $\frac{\lambda\sigma}{n} \to \infty$, θα δούμε ότι με πιθανότητα που τείνει στο 1 , το 0 συμπεριλαμβάνεται στο εφικτό σύνολο για τον Dantzig selector και από την σχέση (6.8) θα ισχύει $\hat{\beta} = 0$.

Θεώρημα 6.1

Υποθέτουμε ότι $\frac{\lambda\sigma}{n} \to \alpha$, $\alpha \in [0, \infty)$, το p είναι σταθερό , το n τείνει στο άπειρο και οι συνθήκες κανονικότητας που αναλύθηκαν πιο πάνω ισχύουν. Εάν το Σ δεν είναι παράλληλο με την l_1-μπάλα, τότε θα έχουμε $\hat{\beta} \xrightarrow{P} \beta_0$, όπου το β_0 αποτελεί λύση του παρακάτω:

$$min_{\beta_0 \in R^p}(\|\beta_0\|_1) \text{ τέτοιο ώστε } \|\Sigma(\beta - \beta_0)\|_\infty \leq \alpha \qquad (6.9)$$

■

Από το Θεώρημα 6.1 προκύπτει ότι όταν το $\alpha = 0$, ισχύει το ενδεχόμενο $\hat{\beta} \to \beta_0$, γεγονός που σημαίνει ότι ο Dantzig selector είναι συνεπής στην εκτίμηση. Από την άλλη σε περίπτωση που έχουμε ότι $0 < \alpha < \infty$, τότε $\|\beta_0\|_1 < \|\beta\|_1$, πράγμα που σημαίνει ότι ο Dantzig selector δεν είναι συνεπής στην εκτίμηση.

6.4 Ο Adaptive Dantzig selector και οι ασυμπτωτικές του ιδιότητες

Σύμφωνα με το Θεώρημα 6.1, διαπιστώνουμε ότι ο Dantzig selector μπορεί να μην είναι συνεπής στα μερικώς γραμμικά μοντέλα. Για τον λόγο αυτό χρησιμοποιούμε τον adaptive Dantzig selector των Dicker και Lin (2009). Ο adaptive Dantzig selector είναι μια γενίκευση του Dantzig selector όπου οι συντελεστές βαρύτητας (*data-dependent weights*) $w_1, \dots, w_p$ εισάγονται στον περιορισμό του αρχικού προβλήματος βελτιστοποίησης, δηλαδή:

$$\begin{gathered} min\left(\sum_{j=1}^{p} w_j|\beta_j|\right) \text{ τέτοιο ώστε} \\ \left|\left(\hat{X}'\left(\hat{Y}-\hat{X}\beta\right)\right)_j\right| \le w_j\lambda\sigma \quad , j=1,..,p \end{gathered} \tag{6.10}$$

Η λύση βελτιστοποίησης στο (6.10) είναι απλά ο εκτιμητής Dantzig selector, ο οποίος εκτελεί ταυτόχρονα την επιλογή μεταβλητών και την εκτίμηση συντελεστών.

Έστω ότι $\beta^0 = W\beta$ και $Z = \hat{X}W^{-1}$, όπου $W = diag(w_1, \dots, w_p)$. Τότε το πρόβλημα βελτιστοποίησης του adaptive Dantzig selector (6.10) μπορεί να αναδιατυπωθεί ως εξής:

$$\begin{gathered} min_{\beta^0 \in R^p}(\|\beta^0\|_1) \\ \text{δεδομένου ότι } \left\|Z'\left(\hat{Y}-Z\,\beta^0\right)\right\|_\infty \le \lambda\cdot\sigma \end{gathered} \tag{6.11}$$

Κατά συνέπεια ο adaptive Dantzig selector αποτελεί μια περίπτωση του Dantzig selector και μπορεί να εφαρμοστεί με την βοήθεια του αλγόριθμου DASSO που προτείνεται από τους James, Radchenkob , και Lv (2009). Σχετικά με την πεπερασμένη ανάλυση του δείγματος μπορούμε να επιλέξουμε την παράμετρο tuning λ με χρήση μιας εκ των μεθόδων CV, GCV και BIC.

Ένα άλλο σημαντικό θέμα σε αυτή την παράγραφο είναι η επιλογή των βαρών (*weights*) που όπως γίνεται αντιληπτό είναι κρίσιμη για την απόδοση του

adaptive Dantzig selector. Σύμφωνα με τους Dicker και Lin (2009), τα βάρη w_j θα πρέπει να είναι αντιστρόφως ανάλογα με τα β_j. Τα λογικά βάρη θα δίνονται από την σχέση $w_j = \left|\hat{\beta}_j(PLS)\right|^{-q}$, όπου $\hat{\beta}_j(PLS)$ είναι η j-οστή συνιστώσα της εκτίμησης των ελάχιστων τετραγώνων του β που ορίζεται ως εξής $\hat{\beta}_j(PLS) = (\hat{X}'\hat{X})^{-1}\hat{X}'\hat{Y}$, $q > 0$. Γενικότερα, τα βάρη μπορούν να ληφθούν ως $w_j = f(|\hat{\beta}^0|)$, όπου $f(\cdot)$ είναι μια θετική φθίνουσα συνάρτηση με $f(0) = \infty$ και $\hat{\beta}^0$ είναι ένας $\sqrt{n}$-συνεπής εκτιμητής του β.

Θεώρημα 6.2

Ασυμπτωτικές ιδιότητες του adaptive Dantzig selector

Υποθέτουμε ότι ισχύουν οι συνθήκες κανονικότητας που δόθηκαν πιο πάνω και ότι $\frac{\lambda\sigma}{\sqrt{n}} w_j \xrightarrow{P} \infty$ για $j \nexists C^*$, $w_j \lambda\sigma = o_p(\sqrt{n})$ για $j \in C^*$, το p είναι σταθερό και το n τείνει στο άπειρο. Τότε ο εκτιμητής adaptive Dantzig selector $\hat{\beta}$ είναι συνεπής για την επιλογή του μοντέλου και $\sqrt{n}(\hat{\beta}_{C^*} - \beta_{C^*}) \xrightarrow{D} N(0, \sigma^2 \Sigma_{C^*,C^*}^{-1})$.

■

Από το Θεώρημα 6.2 συμπεραίνουμε ότι ο adaptive Dantzig selector υιοθετεί τις ιδιότητες oracle. Απαιτεί τον καθορισμό του p να είναι σταθερό, το οποίο είναι κρίσιμο σημείο για την μοναδικότητα του Dantzig selector ακόμη και για τα γραμμικά μοντέλα .

ΚΕΦΑΛΑΙΟ 7

ΕΦΑΡΜΟΓΗ ΠΟΙΝΙΚΟΠΟΙΗΜΕΝΩΝ ΜΕΘΟΔΩΝ

7.1 Εφαρμογή σε Πίνακα Σχεδιασμού Κανονικής Κατανομής

Σε αυτή την ενότητα με τη βοήθεια της γλώσσας προγραμματισμού R καθώς και του στατιστικού πακέτου “flare” (*Family of LASSO Regression*) θα εφαρμόσουμε τις μεθόδους LASSO, Square-root LASSO και Dantzig Selector. Η εφαρμογή αφορά ένα πίνακα σχεδιασμού X με $n = 72$ γραμμές και $d = 256$ στήλες ο οποίος αποτελείται από στοιχεία που προέρχονται από την κανονική κατανομή $N(0,1)$.

Αρχικά θα κατεβάσουμε στην R το πακέτο «flare» που περιέχει τις συναρτήσεις που θα χρησιμοποιήσουμε. Έπειτα εισάγουμε στην R τον $72x256$ πίνακα σχεδιασμού X, το διάνυσμα της παραμέτρου b, το διάνυσμα του σφάλματος το οποίο θα ορίσουμε να ακολουθεί κανονική κατανομή καθώς και το διάνυσμα Y το οποίο θα δίνεται από τον γνωστό τύπο μοντέλου $Y = Xb + \varepsilon$.

Να σημειωθεί εδώ ότι αναλυτικά τα αποτελέσματα υπάρχουν στο τέλος του Κεφαλαίου 7 και στην συγκεκριμένη παράγραφο θα παρουσιάσουμε τα πιο σημαντικά.

Στην κονσόλα της R γράφουμε την ακόλουθη εντολή που αφορά την μέθοδο Dantzig selector:

```
out1=slim(X=X,Y=Y,nlambda=100,lambda.min.ratio=0.3,method="dantzig")
```

Η συνάρτηση slim (*Sparse Linear Regression using Nonsmooth Loss Functions and L1 Regularization*) του πακέτου “flare” χρησιμεύει για την εκτίμηση αραιών

γραμμικών μοντέλων υψηλών διαστάσεων με μεθόδους που ανήκουν στην οικογένεια της LASSO · μεταξύ των οποίων είναι Dantzig Selector, LAD Lasso, Square-root Lasso, Lq Lasso. Σχετικά με τις παραμέτρους της, όπου “nlambda” είναι ο αριθμός των τιμών της παραμέτρου ποινής λ που θέλουμε να μας εμφανίσει. Η παράμετρος “lambda.min.ratio” αντιστοιχεί στην μικρότερη τιμή του λ, ως κλάσμα του άνω φράγματος (lambda.max) της παραμέτρου κανονικοποιήσης (π.χ η μικρότερη τιμή για την οποία όλοι οι συντελεστές είναι μηδέν).

Από την εκτέλεση της παραπάνω εντολής προκύπτει:

[1] 8.57 8.47 8.37 8.26 8.16 8.07 7.97 7.87 7.78 7.68 7.59 7.50 7.41 7.32 7.23

[16] 7.14 7.06 6.97 6.89 6.80 6.72 6.64 6.56 6.48 6.40 6.32 6.25 6.17 6.10 6.02

[31] 5.95 5.88 5.81 5.74 5.67 5.60 5.53 5.47 5.40 5.33 5.27 5.21 5.14 5.08 5.02

[46] 4.96 4.90 4.84 4.78 4.72 4.67 4.61 4.55 4.50 4.44 4.39 4.34 4.29 4.23 4.18

[61] 4.13 4.08 4.03 3.98 3.94 3.89 3.84 3.79 3.75 3.70 3.66 3.61 3.57 3.53 3.49

[76] 3.44 3.40 3.36 3.32 3.28 3.24 3.20 3.16 3.12 3.09 3.05 3.01 2.98 2.94 2.90

[91] 2.87 2.83 2.80 2.77 2.73 2.70 2.67 2.63 2.60 2.57

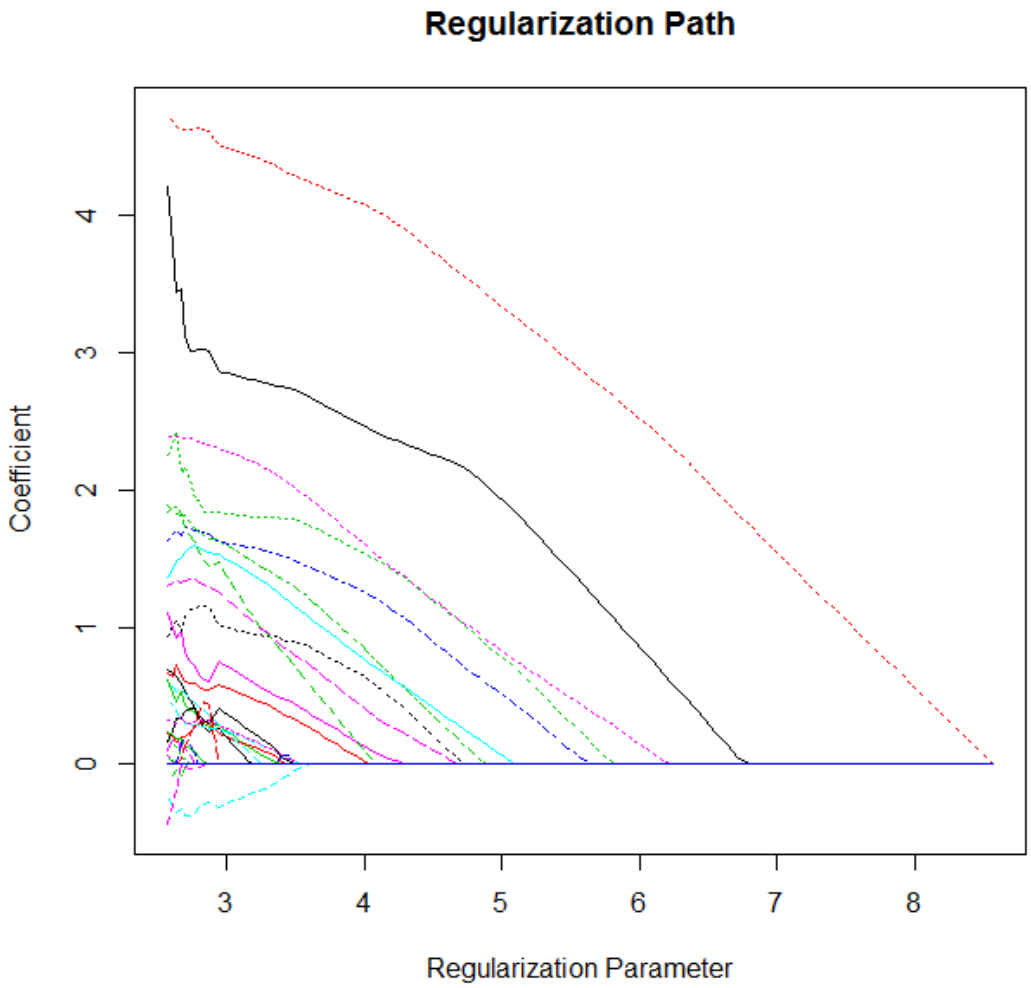
Method = dantzig

Degree of freedom: 0 -----> 28

Runtime: 53.71201 secs

Όπως φαίνεται παρουσιάζονται οι 100 τιμές που θα πάρει η παράμετρος ποινής λ, ο βαθμός ελευθερίας (*Degree of freedom*) που αντιστοιχεί στον αριθμό των μη μηδενικών μεταβλητών , καθώς και ο χρόνος (*runtime*) που απαιτήθηκε για να υπολογιστούν όλα τα πιο πάνω.

Μπορούμε επίσης να απεικονίσουμε τα δεδομένα χρησιμοποιώντας την εντολή plot(out1).



Σχήμα 7.1: Αναπαράσταση της πορείας των συντελεστών σε σχέση με την παράμετρο κανονικοποίησης για την μέθοδο Dantzig Selector.

Σχετικά με την γραφική παράσταση, κάθε γραμμή αντιστοιχεί σε μια μεταβλητή. Δείχνει την πορεία των συντελεστών σε σχέση με την παράμετρο κανονικοποίησης όλου του διανύσματος συντελεστών καθώς το λ παίρνει τις διάφορες τιμές του. Ο άξονας δηλώνει τον αριθμό των μη μηδενικών συντελεστών για το κάθε λ, που είναι οι αποτελεσματικοί βαθμοί ελευθερίας για τον Dantzig Selector.

Ομοίως πράττουμε το ίδιο για την LASSO και την Square-root LASSO και τα αποτελέσματα που προκύπτουν θα είναι αντίστοιχα:

Για την LASSO

[1] 8.570 8.280 8.000 7.730 7.460 7.210 6.960 6.730 6.500 6.280 6.060 5.850

[13] 5.660 5.460 5.280 5.100 4.920 4.760 4.590 4.440 4.290 4.140 4.000 3.860

[25] 3.730 3.600 3.480 3.360 3.250 3.140 3.030 2.930 2.830 2.730 2.640 2.550

[37] 2.460 2.380 2.300 2.220 2.140 2.070 2.000 1.930 1.870 1.800 1.740 1.680

[49] 1.620 1.570 1.520 1.460 1.410 1.370 1.320 1.270 1.230 1.190 1.150 1.110

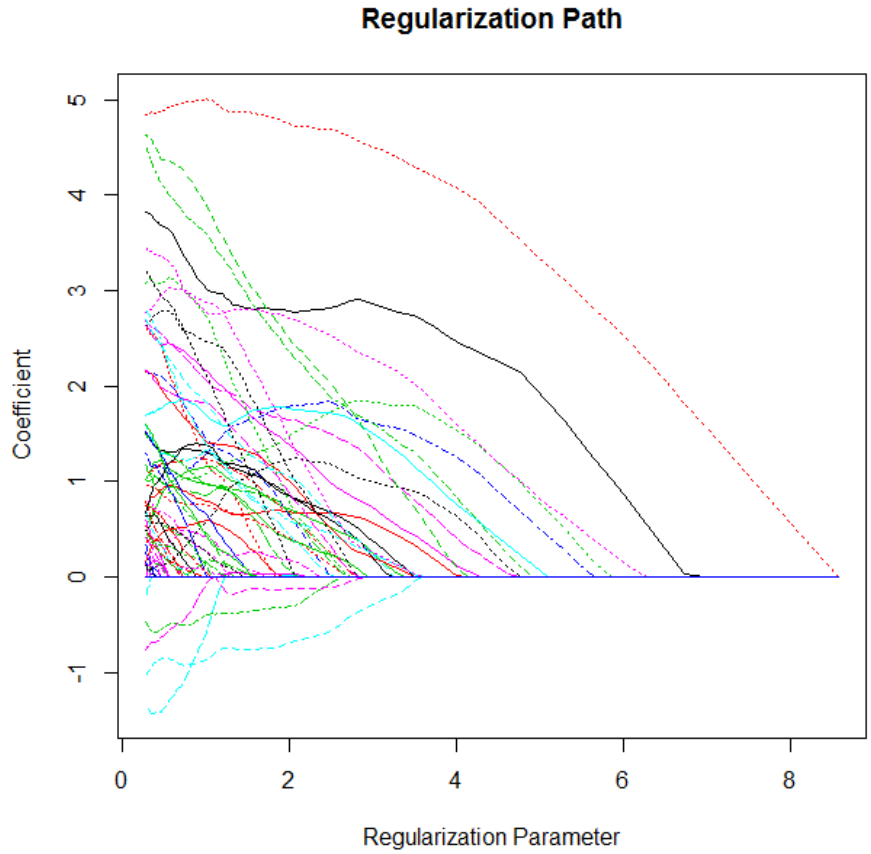
[61] 1.070 1.040 1.000 0.966 0.933 0.901 0.871 0.841 0.812 0.785 0.758 0.732

[73] 0.707 0.683 0.660 0.637 0.616 0.595 0.575 0.555 0.536 0.518 0.500 0.483

[85] 0.467 0.451 0.435 0.421 0.406 0.392 0.379 0.366 0.354 0.342 0.330 0.319

[97] 0.308 0.297 0.287 0.278, Method = lasso , Degree of freedom: 0 -----> 64

Runtime:1.634372 secs



Σχήμα 7.2: Αναπαράσταση της πορείας των συντελεστών σε σχέση με την παράμετρο κανονικοποίησης για την μέθοδο LASSO.

Για την Square-root LASSO

[1] 0.431 0.429 0.427 0.425 0.423 0.421 0.420 0.418 0.416 0.414 0.412 0.410

[13] 0.408 0.407 0.405 0.403 0.401 0.400 0.398 0.396 0.394 0.392 0.391 0.389

[25] 0.387 0.386 0.384 0.382 0.380 0.379 0.377 0.375 0.374 0.372 0.370 0.369

[37] 0.367 0.366 0.364 0.362 0.361 0.359 0.358 0.356 0.354 0.353 0.351 0.350

[49] 0.348 0.347 0.345 0.343 0.342 0.340 0.339 0.337 0.336 0.334 0.333 0.331

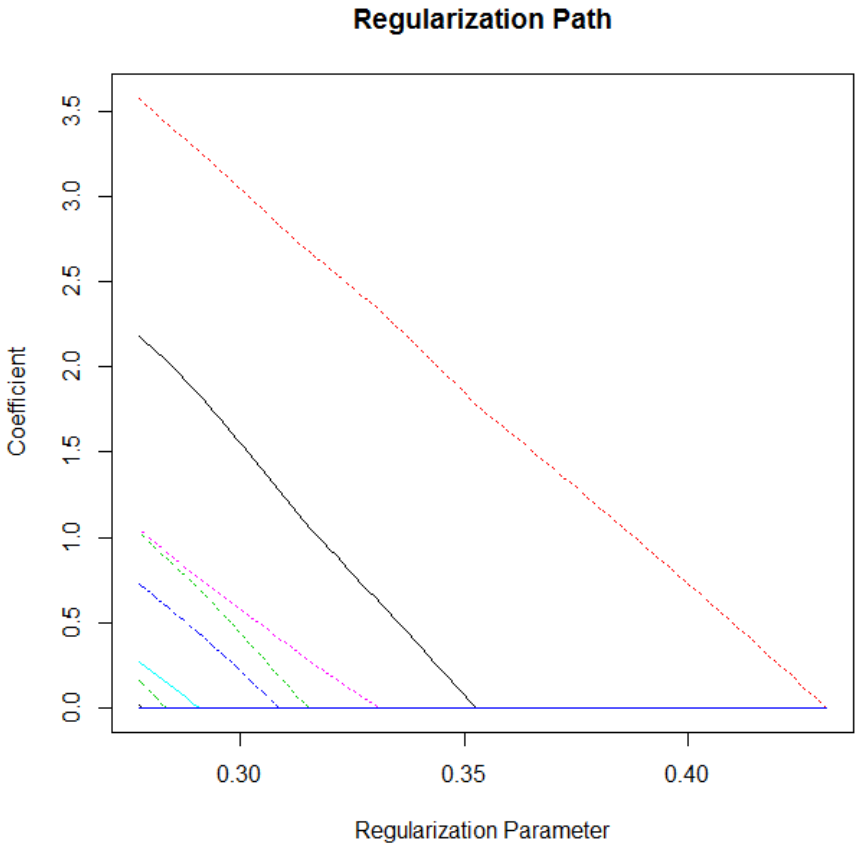
[61] 0.330 0.329 0.327 0.326 0.324 0.323 0.321 0.320 0.319 0.317 0.316 0.314

[73] 0.313 0.312 0.310 0.309 0.307 0.306 0.305 0.303 0.302 0.301 0.299 0.298

[85] 0.297 0.295 0.294 0.293 0.291 0.290 0.289 0.288 0.286 0.285 0.284 0.282

[97] 0.281 0.280 0.279 0.278

Method = lq ,q = 2 loss, SQRT Lasso, Degree of freedom: 1 -----> 8 ,Runtime: 7.080584 secs



Σχήμα 7.3: Αναπαράσταση της πορείας των συντελεστών σε σχέση με την παράμετρο κανονικοποίησης για την μέθοδο Square-root LASSO.

Παρατήρηση 8.1

Αυτό που κάνει ιδιαίτερη εντύπωση στα αποτελέσματα είναι ότι η πιο χρονοβόρα μέθοδος είναι αυτή του Dantzig Selector (*53.71201 secs*) σε σχέση με τις άλλες δύο, ενώ η πιο γρήγορη μέθοδος είναι αυτή της LASSO (*1.634372 secs*). Επίσης παρατηρούμε ότι η μέθοδος LASSO έχει το πιο ευρύ φάσμα τιμών λ αφού παίρνει από πολύ μεγάλες τιμές μέχρι και μικρές ([8.570,0,278]) σε αντίθεση με τον Dantzig selector που παίρνει μόνο μεγαλύτερες τιμές ([8.570,2.570]) και την Square-root LASSO που περιορίζεται μόνο σε μικρότερες τιμές ([0.431,0.278]). Συγχρόνως είναι σημαντικό να τονισθεί ότι η μέθοδος LASSO μπορεί να έχει μέχρι και 64 μη μηδενικές μεταβλητές εκ των 256 που είναι και η διάσταση του συγκεκριμένου μοντέλου ενώ η Dantzig Selector περιορίζεται σε 28 και η Square-root LASSO σε 8.

7.2 Εντολές στην R

```
> library(flare)
Loading required package: lattice
Loading required package: MASS
Loading required package: Matrix
Loading required package: igraph
Attaching package: ‘igraph’
The following objects are masked from ‘package:stats’:
   decompose, spectrum
The following object is masked from ‘package:base’:
   union
```

```
> n=72
> d=256
> X=matrix(rnorm(n*d,mean=0,sd=1),n,d)
> beta=c(1,1,3,5,2,4,rep(1,d-6))
```

> eps=rnorm(n)

> Y=X%*%beta+eps

> out1=slim(X=X,Y=Y,nlambda=100,lambda.min.ratio=0.3,method="dantzig")
//Εδώ γίνεται εφαρμογή της μεθόδου Dantzig Selector

Sparse Linear Regression with L1 Regularization.

Dantzig selector with screening.

slim options summary:

100 lambdas used:

[1] 8.57 8.47 8.37 8.26 8.16 8.07 7.97 7.87 7.78 7.68 7.59 7.50 7.41 7.32 7.23

[16] 7.14 7.06 6.97 6.89 6.80 6.72 6.64 6.56 6.48 6.40 6.32 6.25 6.17 6.10 6.02

[31] 5.95 5.88 5.81 5.74 5.67 5.60 5.53 5.47 5.40 5.33 5.27 5.21 5.14 5.08 5.02

[46] 4.96 4.90 4.84 4.78 4.72 4.67 4.61 4.55 4.50 4.44 4.39 4.34 4.29 4.23 4.18

[61] 4.13 4.08 4.03 3.98 3.94 3.89 3.84 3.79 3.75 3.70 3.66 3.61 3.57 3.53 3.49

[76] 3.44 3.40 3.36 3.32 3.28 3.24 3.20 3.16 3.12 3.09 3.05 3.01 2.98 2.94 2.90

[91] 2.87 2.83 2.80 2.77 2.73 2.70 2.67 2.63 2.60 2.57

Method = dantzig

Degree of freedom: 0 -----> 28

Runtime: 53.71201 secs

>plot(out1)

> out2=slim(X=X,Y=Y,nlambda=100,lambda.min.ratio=0.3,method="lasso")
// Εδώ γίνεται εφαρμογή της μεθόδου LASSO

Sparse Linear Regression with L1 Regularization.

Lasso with screening.

slim options summary:

100 lambdas used:

[1] 8.570 8.280 8.000 7.730 7.460 7.210 6.960 6.730 6.500 6.280 6.060 5.850

[13] 5.660 5.460 5.280 5.100 4.920 4.760 4.590 4.440 4.290 4.140 4.000 3.860

[25] 3.730 3.600 3.480 3.360 3.250 3.140 3.030 2.930 2.830 2.730 2.640 2.550

[37] 2.460 2.380 2.300 2.220 2.140 2.070 2.000 1.930 1.870 1.800 1.740 1.680

[49] 1.620 1.570 1.520 1.460 1.410 1.370 1.320 1.270 1.230 1.190 1.150 1.110

[61] 1.070 1.040 1.000 0.966 0.933 0.901 0.871 0.841 0.812 0.785 0.758 0.732

[73] 0.707 0.683 0.660 0.637 0.616 0.595 0.575 0.555 0.536 0.518 0.500 0.483

[85] 0.467 0.451 0.435 0.421 0.406 0.392 0.379 0.366 0.354 0.342 0.330 0.319

[97] 0.308 0.297 0.287 0.278

Method = lasso

Degree of freedom: 0 -----> 64

Runtime: 1.634372 secs

>plot(out2)

> out3=slim(X=X,Y=Y,nlambda=100,lambda.min.ratio=0.3,method="lq")
// Εδώ γίνεται εφαρμογή της μεθόδου Square-root LASSO

Sparse Linear Regression with L1 Regularization.

Square root Lasso with screening.

slim options summary:

100 lambdas used:

[1] 0.431 0.429 0.427 0.425 0.423 0.421 0.420 0.418 0.416 0.414 0.412 0.410

[13] 0.408 0.407 0.405 0.403 0.401 0.400 0.398 0.396 0.394 0.392 0.391 0.389

[25] 0.387 0.386 0.384 0.382 0.380 0.379 0.377 0.375 0.374 0.372 0.370 0.369

[37] 0.367 0.366 0.364 0.362 0.361 0.359 0.358 0.356 0.354 0.353 0.351 0.350

[49] 0.348 0.347 0.345 0.343 0.342 0.340 0.339 0.337 0.336 0.334 0.333 0.331

[61] 0.330 0.329 0.327 0.326 0.324 0.323 0.321 0.320 0.319 0.317 0.316 0.314

[73] 0.313 0.312 0.310 0.309 0.307 0.306 0.305 0.303 0.302 0.301 0.299 0.298

[85] 0.297 0.295 0.294 0.293 0.291 0.290 0.289 0.288 0.286 0.285 0.284 0.282

[97] 0.281 0.280 0.279 0.278

Method = lq

q = 2 loss, SQRT Lasso

Degree of freedom: 1 -----> 8

Runtime: 7.080584 secs

>plot(out3)

ΠΑΡΑΡΤΗΜΑ Α

Α.1 Αποδείξεις για το Κεφάλαιο 5

Σε αυτό το σημείο θα αποδειχθούν τα τρία θεωρήματα που παρουσιάστηκαν στο Κεφάλαιο 5. Έστω ότι $X^1, \dots, X^p \in R^n$ είναι οι p στήλες τους πίνακα X έτσι ώστε $X\beta = \beta_1 X^1 + \cdots + \beta_p X^p$ και $(X^* y)_j = \langle y, X^j \rangle, 1 \le j \le p$. Υπενθυμίζουμε ότι οι στήλες του πίνακα X είναι κανονικοποιημένες για να έχουν μοναδιαία νόρμα, δηλαδή να ισχύει $\|X^j\|_{\ell_2} = 1$.

Να σημειωθεί ότι αρκεί να αποδείξουμε τα Θεωρήματα μας για $\sigma = 1$, δεδομένου ότι προκύπτουν εύκολα και οι αποδείξεις για τις γενικές περιπτώσεις. Ως εκ τούτου, από τώρα και στο εξής υποθέτουμε ότι $\sigma = 1$. Μια σημαντική παρατήρηση είναι ότι με μεγάλη πιθανότητα , το $z \sim N(0, I_n)$ θα υπακούει στην συνθήκη ορθογωνιότητας (*orthogonality condtition*)

$$\left|\langle z, X^j \rangle\right| \le \lambda_p \text{ για όλα τα } 1 \le j \le p \qquad (A.1)$$

για $\lambda_p = \sqrt{2\log p}$. Αυτό είναι κάτι που ισχύει γενικά και προκύπτει από το γεγονός ότι για κάθε j, $Z_j := \langle z, X^j \rangle \sim N(0,1)$. Θα δούμε ότι εάν ισχύει η (Α.1) τότε ισχύει και η ανισότητα (5.9) του Θεωρήματος 5.1. Να σημειωθεί ότι για κάθε $u > 0, P\left(sup_j |Z_j| > u\right) \le 2p \cdot \phi(u)/u$, όπου $\phi(u) := (2\pi)^{-1/2} e^{-u^2/2}$. Όπως προαναφέρθηκε, στην περίπτωση που οι στήλες δεν έχουν μοναδιαία νόρμα τότε κάποιος θα λάμβανε το ίδιο αποτέλεσμα για $\lambda_p = \sqrt{1+\delta_1} \cdot \sqrt{2\log p}$ δεδομένου ότι $\|X^j\|_{\ell_2} \le \sqrt{1+\delta_1}$.

Α.1.2 Γεωμετρία υψηλών διαστάσεων

Πριν προχωρήσουμε στην απόδειξη των Θεωρημάτων θα γίνει μια αναφορά στην γεωμετρία των υψηλών διαστάσεων (*high dimensional geometry*) μιας και

θα χρειαστεί στις μετέπειτα αποδείξεις. Ας επικεντρωθούμε αρχικά στο Θεώρημα 5.1 και να υποθέσουμε ότι έχουμε το μοντέλο $y = X\beta + z$, όπου τα z υπακούνε στην συνθήκη ορθογωνιότητας (A.1) για κάποια λ_p. Έστω ότι το $\hat{\beta}$ ελαχιστοποιεί το πρόβλημα (5.7). Σαφώς, το διάνυσμα των πραγματικών παραμέτρων β θα είναι εφικτό και επομένως θα ισχύει

$$\|\hat{\beta}\|_{\ell_1} \leq \|\beta\|_{\ell_1}$$

Αναλύεται το $\hat{\beta}$ ως $\hat{\beta} = \beta + h$ και ορίζουμε το T_0 να είναι ο φορέας του β, δηλαδή $T_0 = \{i : \beta_i \neq 0\}$. Τότε το h θα υπακούει σε δυο γεωμετρικούς περιορισμούς.

1. Πρώτα σύμφωνα με τους Donoho και Huo (2001),

$$\|\beta\|_{\ell_1} - \|h_{T_0}\|_{\ell_1} + \|h_{T_0^c}\|_{\ell_1} \leq \|\beta + h\|_{\ell_1} \leq \|\beta\|_{\ell_1}$$

όπου η i-οστή συνιστώσα του διανύσματος h_{T_0}είναι αυτή του h εάν το i $\in T_0$ ενώ θα είναι μηδέν σε οποιαδήποτε άλλη περίπτωση. Έτσι το h θα υπακούει στον περιορισμό του κώνου (*cone constraint*)

$$\|h_{T_0^c}\|_{\ell_1} \leq \|h_{T_0}\|_{\ell_1} \qquad (A.2)$$

2. Έπειτα, αφού ισχύει

$$\langle z - r, X^j \rangle = \langle X\hat{\beta} - X\beta, X^j \rangle = \langle Xh, X^j \rangle$$

τότε από την τριγωνική ανισότητα θα προκύπτει

$$\|X^* Xh\|_{\ell_\infty} \leq 2\lambda_P \qquad (A.3)$$

Το αξιοσημείωτο με αυτούς τους δύο γεωμετρικούς περιορισμούς είναι ότι, όπως θα δούμε και στην συνέχεια, δείχνουν ότι το h είναι μικρό στην ℓ_2-νόρμα.

Με λίγα λόγια, θα δείξουμε ότι η σχέση

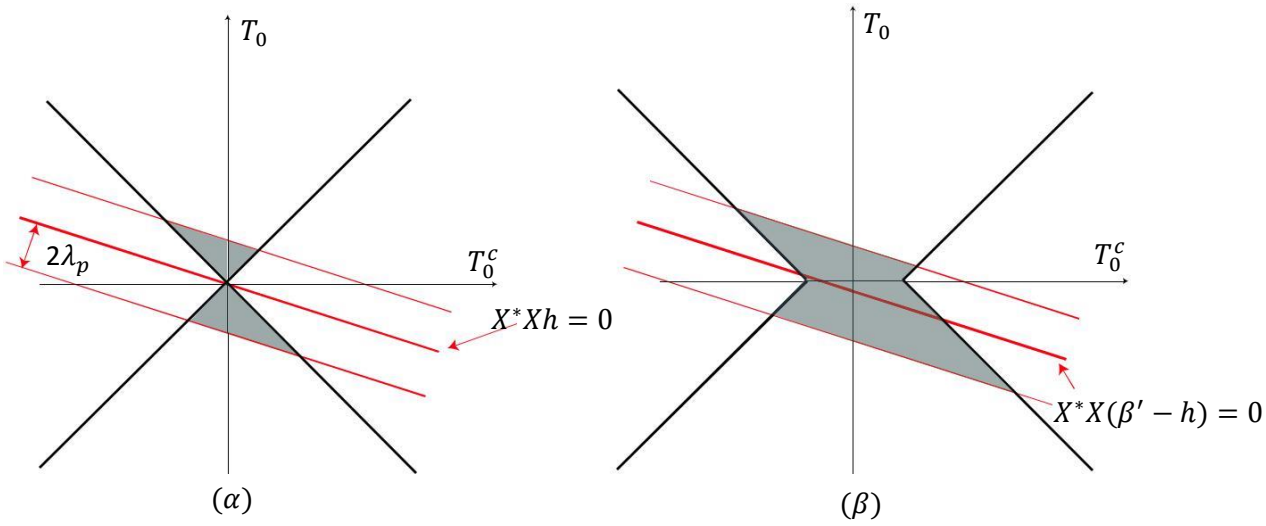
$$\sup_{h\in R^p} \|h\|_{\ell_2}^2 \text{ δεδομένου ότι}$$
$$\left\|h_{T_0^c}\right\|_{\ell_1} \le \left\|h_{T_0}\right\|_{\ell_1} \text{ και } \|X^*Xh\|_{\ell_\infty} \le 2\lambda_p \qquad (A.4)$$

υπακούει στο επιθυμητό φράγμα, δηλαδή $O\left(\lambda_p^2 \cdot |T_0|\right)$. Αυτό απεικονίζεται στο Σχήμα Α.1 (α). Συνεπώς, το στατιστικό μας ερώτημα είναι βαθιά συνδεδεμένο με την γεωμετρία των χώρων Banach σε υψηλές διαστάσεις, και γενικά με οποιουσδήποτε χώρους υψηλών διαστάσεων. Για την γενική περίπτωση, εξετάζουμε το σύνολο των δεικτών $T_0 := \{i : |\beta_i| > \sigma\}$ και ορίζουμε το β_{T_0} να είναι το διάνυσμα που ισοδυναμεί με το β στο T_0 και να μηδενίζεται σε κάθε άλλη περίπτωση· δηλαδή $\beta = \beta_{T_0} + \beta_{T_0^c}$.

Υποθέτουμε τώρα ότι εάν τα β_{T_0} είναι εφικτά , τότε θα έχυμε $\left\|\hat{\beta}\right\|_{\ell_1} \le \left\|\beta_{T_0}\right\|_{\ell_1}$ και γράφοντας $\hat{\beta} = \beta_{T_0} + h$ θα προκύπτει

$$\left\|\hat{\beta} - \beta_{T_0}\right\|_{\ell_2}^2 = O(\log p) \cdot |T_0| \cdot \sigma^2$$

Από την σχέση $\left\|\hat{\beta} - \beta\right\|_{\ell_2}^2 \le 2\left\|\hat{\beta} - \beta_{T_0}\right\|_{\ell_2}^2 + 2\left\|\beta - \beta_{T_0}\right\|_{\ell 2}^2$, με αντικατάσταση θα έχουμε $\left\|\hat{\beta} - \beta\right\|_{\ell_2}^2 = O(\log p) \cdot |T_0| \cdot \sigma^2 + 2\sum_{i:|\beta_i|<\sigma} \beta_i^2$ · που είναι το περιεχόμενο της (5.23). Παρόλα αυτά, ενώ το β_{T_0} μπορεί να είναι εφικτό για πολλά S-αραιά διανύσματα β , εντούτοις για μερικά δεν είναι και εκεί τα πράγματα γίνονται πιο περίπλοκα.



Σχήματα Α.1 (α) , Α.1 (β): Τα σχήματα απεικονίζουν την γεωμετρία των περιορισμών. Στο Σχήμα (α), η σκιασμένη περιοχή αντιπροσωπεύει το σύνολο των h που υπακούει τόσο την σχέση (Α.2) και την (Α.3) . Το σχήμα (β) απεικονίζει την κατάσταση σε πιο γενική περίπτωση.

Α.1.2 Απόδειξη Θεωρήματος 5.1

Λήμμα Α.1

Ας υποθέσουμε ότι T_0 είναι ένα σύνολο με πληθικότητα S όπου $\delta + \theta < 1$. Για το διάνυσμα $h \in R^p$, ορίζουμε T_1 να είναι οι S μεγαλύτερες θέσεις του h έξω από το T_0. Ορίζουμε $T_{01} = T_0 \cup T_1$. Τότε

$$\|h\|_{\ell_2(T_{01})} \leq \frac{1}{1-\delta}\left\|X_{T_{01}}^T X h\right\|_{\ell_2} + \frac{\theta}{(1-\delta)S^{1/2}}\|h\|_{\ell_1(T_0^c)} \text{ και}$$

$$\|h\|_{\ell_2}^2 \leq \|h\|_{\ell_2(T_{01})}^2 + S^{-1}\|h\|_{\ell_1(T_0^c)}^2$$

■

Απόδειξη Λήμματος Α.1

Θα εξετάσουμε τον περιορισμένο μετασχηματισμό $X_{T_{01}} : R^{T_{01}} \to R^n, X_{T_{01}} c := \sum_{j \in T_{01}} c_j X^j$. Έστω $V \subset R^n$ να είναι η θήκη του $\{X^j : j \in T_{01}\}$. Τότε το V είναι προφανές το πεδίο τιμών (εικόνα) του $X_{T_{01}}$ και επίσης το ορθογώνιο συμπλήρωμα του πυρήνα του $X_{T_{01}}^T$ όπου λέει ότι R^n είναι το ορθογώνιο άθροισμα $V \oplus V^\perp$. Επειδή $\delta < 1$ γνωρίζουμε ότι ο τελεστής $X_{T_{01}}$ είναι αμφιμονοσήμαντος από το $R^{T_{01}}$ στο V με ιδιάζουσες τιμές μεταξύ $\sqrt{1-\delta}$ και $\sqrt{1+\delta}$. Συνεπώς, για κάθε $c \in \ell_2(T_{01})$ θα έχουμε

$$\sqrt{1-\delta}\|c\|_{\ell_2} \leq \left\|X_{T_{01}} c\right\|_{\ell_2} \leq \sqrt{1+\delta}\|c\|_{\ell_2}$$

Επίσης, ορίζουμε P_V να είναι η ορθογώνια προβολή μέσα στο V, και θα έχουμε $X_{T_{01}}^T w = X_{T_{01}}^T P_V w$ για κάθε $w \in R^n$ και άρα θα προκύπτει ότι

$$\sqrt{1-\delta}\|P_V w\|_{\ell_2} \leq \left\|X_{T_{01}}^T w\right\|_{\ell_2} \leq \sqrt{1+\delta}\|P_V w\|_{\ell_2} \qquad (A.5)$$

Εφαρμόζουμε όπου $w := Xh$ και καταλήγουμε στη σχέση

$$\|P_V Xh\|_{\ell_2} \leq (1-\delta)^{-1/2} \left\| X_{T_{01}}^T Xh \right\|_{\ell_2} \qquad (A.6)$$

Το επόμενο βήμα είναι να εξάγουμε ένα χαμηλότερο φράγμα στο $P_V Xh$. Για να γίνει αυτό ξεκινάμε με την διαίρεση του T_0^c σε υποσύνολα μεγέθους S και απαριθμούμε το T_0^c ως $n_1, n_2, \dots, n_{p-|T_0|}$ κατά φθίνουσα τάξη μεγέθους του $h_{T_0^c}$. Ορίζουμε $T_j := \{n_\ell, (j-1)S + 1 \leq \ell \leq jS\}$. Δηλαδή το T_1 είναι όπως πρίν και περιέχει τους δείκτες των S μεγαλύτερων συντελεστών του $h_{T_0^c}$, το T_2 περιέχει τους δείκτες των επόμενων S συντελεστών και ούτω κάθε εξής.

Στη συνέχεια αναλύουμε το $P_V Xh$ ως

$$P_V Xh = P_V Xh_{T_{01}} + \sum_{j \geq 2} P_V Xh_{T_j} \qquad (A.7)$$

Εξ' ορισμού $Xh_{T_{01}} \in V$ και $P_V Xh_{T_{01}} = Xh_{T_{01}}$. Περαιτέρω, δεδομένου ότι το P_V είναι μια ορθογώνια προβολή πάνω στη θήκη του X^j για $j \in T_{01}$, τότε έχουμε $P_V Xh_{T_j} = \sum_{j \in T_{01}} c_j X^j$ για κάποιους συντελεστές c_j και έτσι θα ισχύει η ακόλουθη σχέση

$$\left\| P_V Xh_{T_j} \right\|_{\ell_2}^2 = \langle P_V Xh_{T_j}, Xh_{T_j} \rangle \qquad (A.8)$$

Λόγω της περιορισμένης ορθογωνιότητας που ακολουθείται από περιορισμένη ισομετρία προκύπτει

$$\langle P_V Xh_{T_j}, Xh_{T_j} \rangle \leq \theta \left(\sum_{j \in T_{01}} |c_j|^2 \right)^{1/2} \left\| h_{T_j} \right\|_{\ell_2} \leq \frac{\theta}{\sqrt{1-\delta}} \left\| P_V Xh_{T_j} \right\|_{\ell_2} \left\| h_{T_j} \right\|_{\ell_2}$$

τα οποία σε συνδυασμό με την σχέση (A.8) δίνουν

$$\left\| P_V Xh_{T_j} \right\|_{\ell_2} \leq \frac{\theta}{\sqrt{1-\delta}} \left\| h_{T_j} \right\|_{\ell_2} \qquad (A.9)$$

Έπειτα, αναπτύσσουμε ένα κάτω φράγμα στο $\sum_{j\geq 2}\left\|h_{T_j}\right\|_{\ell_2}$ όπως αυτό του Candès et.al. (2005). Εκ κατασκευής, το μέγεθος κάθε συνιστώσας $h_{T_{j+1}}[i]$ του $h_{T_{j+1}}$ έχει μικρότερο μέγεθος από τη μέση τιμή των μεγεθών των συνιστωσών του h_{T_j} :

$$\left|h_{T_{j+1}}[i]\right| \leq \left\|h_{T_j}\right\|_{\ell_1}/S$$

Τότε $\left\|h_{T_{j+1}}\right\|_{\ell_2}^2 \leq \left\|h_{T_j}\right\|_{\ell_1}^2/S$ και επομένως

$$\sum_{j\geq 2}\left\|h_{T_j}\right\|_{\ell_2} \leq S^{-1/2}\sum_{j\geq 1}\left\|h_{T_j}\right\|_{\ell_1} = S^{-1/2}\|h\|_{\ell_1(T_0^c)} \qquad (A.10)$$

Με λίγα λόγια, το $Xh_{T_{01}}$ υπακούει στη σχέση $\left\|Xh_{T_{01}}\right\|_{\ell_2} \geq \sqrt{1-\delta}\left\|h_{T_{01}}\right\|_{\ell_2}$ από την περιορισμένη ισομετρία και άρα

$$\sum_{j\geq 2}\left\|P_V X h_{T_j}\right\|_{\ell_2} \leq \theta(1-\delta)^{-1/2}S^{-1/2}\|h\|_{\ell_1(T_0^c)} \Rightarrow$$

$$\|P_V Xh\|_{\ell_2} \geq \sqrt{1-\delta}\|h\|_{\ell_2(T_{01})} - \frac{\theta}{\sqrt{1-\delta}}S^{-1/2}\|h\|_{\ell_1(T_0^c)}$$

Συνδυάζοντας την πιο πάνω σχέση με την (A.6) αποδεικνύεται το πρώτο κομμάτι του Λήμματος.

Όσων αφορά το δεύτερο κομμάτι, παρατηρούμε ότι η k-οστή μεγαλύτερη τιμή του $h_{T_0^c}$ υπακούει στη σχέση

$$\left|h_{T_0^c}\right|_{(k)} \leq \left\|h_{T_0^c}\right\|_{\ell_1}/k$$

Και επομένως

$$\left\|h_{T_{01}^c}\right\|_{\ell_2}^2 \leq \left\|h_{T_0^c}\right\|_{\ell_1}^2 \sum_{k\geq S+1} 1/k^2 \leq \left\|h_{T_0^c}\right\|_{\ell_1}^2/S$$

Και έτσι αποδεικνύεται το Λήμμα Α.1 .

■

Η απόδειξη του Θεωρήματος 5.1 προκύπτει εύκολα από το παραπάνω Λήμμα. Παρατηρούμε ότι από την μία μεριά η (Α.2) μας δίνει

$$\left\|h_{T_0^c}\right\|_{\ell_1} \leq \left\|h_{T_0}\right\|_{\ell_1} \leq S^{1/2}\left\|h_{T_0}\right\|_{\ell_2}$$

Ενώ από την άλλη μεριά η σχέση (Α.3) μας δίνει

$$\left\|X_{T_{01}}^T Xh\right\|_{\ell_2} \leq (2S)^{1/2} \cdot 2\lambda_p$$

δεδομένου ότι κάθε μια εκ των $2S$ συνιστωσών του $X_{T_{01}}^T Xh$ είναι το πολύ $2\lambda_p$ όπως είδαμε από την (Α.3).

Τέλος, εφαρμόζουμε το Λήμμα Α.1 και προκύπτει

$$\|h\|_{\ell_2(T_{01})} \leq \frac{1}{1-\delta-\theta} \cdot \sqrt{2S} \cdot 2\lambda_p$$

Έτσι η απόδειξη του Θεωρήματος 5.1 προκύπτει από την παρακάτω σχέση

$$\|h\|_{\ell_2}^2 \leq 2\|h\|_{\ell_2(T_{01})}^2$$

■

Α.1.3 Απόδειξη Θεωρήματος 5.3

Η απόδειξη του Θεωρήματος (5.3) στηρίζεται στην ίδια λογική με αυτή του Θεωρήματος 5.1 . Ορίζουμε το T_0 να είναι το σύνολο των S μεγαλύτερων στοιχείων του β , και αναλύουμε όπως και πριν το $\hat{\beta} = \beta + h$. Εάν το z υπακούει στη συνθήκη ορθογωνιότητας (5.5) , τότε το β θα είναι εφικτό και

$$\left\|\beta_{T_0}\right\|_{\ell_1} - \left\|h_{T_0}\right\|_{\ell_1} + \left\|h_{T_0^c}\right\|_{\ell_1} - \left\|\beta_{T_0^c}\right\|_{\ell_1} \leq \|\beta + h\|_{\ell_1} \leq \|\beta\|_{\ell_1}$$

το οποίο μας δίνει

$$\left\|h_{T_0^c}\right\|_{\ell_1} \leq \left\|h_{T_0}\right\|_{\ell_1} + 2\left\|\beta_{T_0^c}\right\|_{\ell_1}$$

Η παρουσία του επιπλέον όρου είναι η μόνη διαφορά σε σχέση με πριν. Ολοκληρώνουμε συνδυάζοντας το Λήμμα Α.1 και την σχέση (Α.3) όπου

$$\left\|h_{T_0^c}\right\|_{\ell_2} \leq \frac{C}{1-\delta-\theta} \cdot \left(\lambda_p \cdot S^{1/2} + \left\|\beta_{T_0^c}\right\|_{\ell_1} \cdot S^{-1/2}\right)$$

Το δεύτερο μέρος του Λήμματος Α.1 δίνει $\|h\|_{\ell_2} \leq 2\|h\|_{\ell_2(T_{01})} + \left\|\beta_{T_0^c}\right\|_{\ell_1} \cdot S^{-1/2}$. Δεδομένου ότι το β υπακούει στη συνθήκη Decay (5.27), $\left\|\beta_{T_0^c}\right\|_{\ell_1} \cdot S^{-1/2} \leq C \cdot R \cdot S^{-r}$, με $r = 1/s - 1/2$, το διαπιστώνουμε για όλα τα $S \leq S_*$.

Και έτσι προκύπτει το ζητούμενο

$$\left\|h_{T_0^c}\right\|_{\ell_2} \leq \frac{C}{1-\delta_{S_*}-\theta_{\theta_*,2S_*}} \cdot \left(\lambda_p \cdot S^{1/2} + R \cdot S^{-r}\right)$$

■

Α.1.4 Απόδειξη Θεωρήματος 5.2

Λήμμα Α.2

Για κάθε διάνυσμα β, έχουμε ότι

$$\|X\beta\|_{\ell_2} \leq \sqrt{1+\delta}\left(\|\beta\|_{\ell_2} + (2S)^{-1/2}\|\beta\|_{\ell_1}\right)$$

■

Απόδειξη Λήμματος Α.2

Έστω το T_1να είναι οι $2S$ μεγαλύτερες θέσεις του β, τότε το T_2 θα είναι το αμέσως μεγαλύτερο σύνολο και ούτω καθεξής.

Τότε

$$\|X\beta\|_{\ell_2} \leq \left\|X\beta_{T_1}\right\|_{\ell_2} + \sum_{j\geq 2}\left\|X\beta_{T_j}\right\|_{\ell_2}$$

Από την περιορισμένη ισομετρία θα έχουμε

$$\left\|X\beta_{T_1}\right\|_{\ell_2} \leq (1+\delta)^{1/2}\left\|\beta_{T_1}\right\|_{\ell_2} \leq (1+\delta)^{1/2}\|\beta\|_{\ell_2}$$

και

$$\left\|X\beta_{T_j}\right\|_{\ell_2} \leq (1+\delta)^{1/2}\left\|\beta_{T_j}\right\|_{\ell_2} \leq (1+\delta)^{1/2}(2S)^{-1/2}\left\|\beta_{T_{j-1}}\right\|_{\ell_1}$$

■

Ακολουθεί η απόδειξη του Θεωρήματος 5.2 με την βοήθεια των πιο πάνω. Ως συνήθως, θα ορίσουμε $\hat{\beta}$ να είναι ο ℓ_1 ελαχιστοποιητής δεδομένου του ακόλουθου περιορισμού

$$\left\|X^*\left(X\hat{\beta}-y\right)\right\|_{\ell_\infty} = \sup_{1\leq j\leq p}\left|\langle X\hat{\beta}-y, X^j\rangle\right| \leq (1+t^{-1})\lambda \qquad (A.11)$$

όπου $\lambda := \sqrt{2\log p}$.

Χωρίς βλάβη της γενικότητας μπορούμε να διατάξουμε τα β_j κατά φθίνουσα σειρά μεγέθους

$$|\beta_1| \geq |\beta_2| \geq \cdots \geq \left|\beta_p\right| \qquad (A.12)$$

Συγκεκριμένα από την υπόθεση αραιότητας του β έχουμε ότι

$$\beta_j = 0 \text{ για όλα τα } j > S \qquad (A.13)$$

Ειδικότερα παρατηρούμε ότι

$$\sum_j \min\left(\beta_j^2, \lambda^2\right) \leq S\cdot\lambda^2$$

Επιπλέον το S_0 θα είναι ο μικρότερος δεκαδικός έτσι ώστε

$$\sum_j \min\left(\beta_j^2, \lambda^2\right) \leq S_0\cdot\lambda^2 \qquad (A.14)$$

Συνεπώς $0 \le S_0 \le S$ και

$$S_0 \cdot \lambda^2 \le \lambda^2 + \sum_j \min\left(\beta_j^2, \lambda^2\right) \qquad (A.15)$$

Επίσης, παρατηρούμε από την εξίσωση (A.12) ότι

$$S_0 \cdot \lambda^2 \ge \sum_{j=1}^{S_0+1} \min\left(\beta_j^2, \lambda^2\right) \ge (S_0 + 1) \min\left(\beta_{S_0+1}^2, \lambda^2\right)$$

Και ως εκ τούτου το $\min\left(\beta_{S_0+1}^2, \lambda^2\right)$ είναι αυστηρά μικρότερο από το λ^2. Από την (A.12) συμπεραίνουμε ότι

$$\beta_j < \lambda \text{ για όλα τα } j > S_0. \qquad (A.16)$$

Κάνουμε την ανάλυση του β ως $\beta = \beta^{(1)} + \beta^{(2)}$ όπου

$$\beta_j^{(1)} = \beta_j \cdot 1_{1 \le j \le S_0}$$

$$\beta_j^{(2)} = \beta_j \cdot 1_{j > S_0}$$

Επομένως $\beta^{(1)}$ είναι η φραγμένη μορφή του β, που εντοπίζεται στο σύνολο

$$T_0 := \{1, \dots, S_0\}$$

Από την σχέση (A.16), το $\beta^{(2)}$ είναι S-αραιό με

$$\left\|\beta^{(2)}\right\|_{\ell_2}^2 = \sum_{j > S_0} \min\left(\beta_j^2, \lambda^2\right) \le S_0 \cdot \lambda^2$$

Με χρήση του Πορίσματος A.1, που παρατίθεται παρακάτω, επιτρέπεται η α-νάλυση $\beta^{(2)} = \beta' + \beta''$ όπου

$$\|\beta'\|_{\ell_2} \le \frac{1+\delta}{1-\delta-\theta} \lambda \cdot S_0^{1/2} \qquad (A.17)$$

$$\|\beta'\|_{\ell_1} \leq \frac{1+\delta}{1-\delta-\theta}\lambda \cdot S_0 \qquad (A.18)$$

και

$$\|X^*X\beta''\|_{\ell_\infty} < \frac{1+\delta^2}{1-\delta-\theta}\lambda \qquad (A.19)$$

Χρησιμοποιούμε την ανάλυση και παρατηρούμε ότι

$$X^*\left(X\left(\beta^{(1)}+\beta'\right)-y\right) = -X^*X\beta'' - X^*z$$

Συνεπώς από την (A.1), (A.19) προκύπτει

$$\left\|X^*\left(X\left(\beta^{(1)}+\beta'\right)-y\right)\right\|_{\ell_\infty} \leq \left(1+\frac{1-\delta^2}{1-\delta-\theta}\right)\lambda \qquad (A.20)$$

Από την υπόθεση έχουμε $(1-\delta-\theta)^{-1} \leq t^{-1}$ και άρα $\beta^{(1)}+\beta'$ είναι εφικτή· η οποία με τη σειρά της προϋποθέτει ότι

$$\left\|\hat{\beta}\right\|_{\ell_1} \leq \left\|\beta^{(1)}+\beta'\right\|_{\ell_1} \leq \left\|\beta^{(1)}\right\|_{\ell_1} + \frac{(1+\delta)}{1-\delta-\theta}S_0 \cdot \lambda$$

Αναλύουμε το $\hat{\beta} = \beta^{(1)} + h$. Τότε $\left\|\hat{\beta}\right\|_{\ell_1} \geq \left\|\beta^{(1)}\right\|_{\ell_1} - \|h\|_{\ell_1(T_0)} + \|h\|_{\ell_1(T_0^c)}$ έτσι ώστε

$$\|h\|_{\ell_1(T_0^c)} \leq \|h\|_{\ell_1(T_0)} + \frac{(1+\delta)}{1-\delta-\theta}S_0 \cdot \lambda \qquad (A.21)$$

Και από την σχέση (A.11) και (A.20) , καταλήγουμε στο συμπέρασμα ότι

$$\|X^*X(\beta'-h)\|_{\ell_\infty} \leq 2\left(1+\frac{1-\delta^2}{1-\delta-\theta}\right)\lambda \qquad (A.22)$$

Το Σχήμα Α.1 (β) απεικονίζει σχηματικά αυτά τα δύο προβλήματα

Το υπόλοιπο μέρος της απόδειξης είναι ουσιαστικά εκείνο του Θεωρήματος 5.1. Από το Λήμμα Α.1 θα έχουμε

$$\|h_{01}\|_{\ell_2} \le \frac{1}{1-\delta}\left\|X_{T_{01}}^T X h\right\|_{\ell_2} + \frac{\theta}{(1-\delta)S_0^{1/2}}\|h\|_{\ell_1(T_0^c)}$$

Από την σχέση (Α.22) θα έχουμε

$$\left\|X_{T_{01}}^T X(\beta' - h)\right\|_{\ell_2} \le 2\sqrt{2}\left(1 + \frac{1-\delta^2}{1-\delta-\theta}\right)S_0^{1/2}\cdot\lambda$$

Ενώ από το Λήμμα Α.2 και τις σχέσεις (Α.18), (Α.17) προκύπτει

$$\|X\beta'\|_{\ell_2} \le \left(1 + 1/\sqrt{2}\right)\frac{(1+\delta)^{3/2}}{1-\delta-\theta}S_0^{1/2}\cdot\lambda$$

Επομένως από την περιορισμένη ισομετρία οδηγούμαστε στην σχέση

$$\left\|X_{T_{01}}^T X\beta'\right\|_{\ell_2} \le \left(1 + 1/\sqrt{2}\right)\frac{(1+\delta)^{2}}{1-\delta-\theta}S_0^{1/2}\cdot\lambda$$

Με λίγα λόγια ισχύει

$$\left\|X_{T_{01}}^T X h\right\|_{\ell_2} \le C_0\cdot S_0^{1/2}\cdot\lambda$$

όπου C_0 ορίστηκε στη σχέση (5.24). Έτσι συμπεραίνουμε ότι

$$\|h_{01}\|_{\ell_2} \le \frac{C_0}{1-\delta}S_0^{1/2}\cdot\lambda + \frac{\theta}{(1-\delta)S_0^{1/2}}\|h\|_{\ell_1(T_0^c)}$$

Τέλος, το φράγμα (Α.21) μας δίνει

$$\|h\|_{\ell_1(T_0^c)} \le S_0^{1/2}\|h_{01}\|_{\ell_2} + \frac{1+\delta}{1-\delta-\theta}S_0\cdot\lambda$$

Και επομένως

$$\|h_{01}\|_{\ell_2} \le C'_0\cdot S_0^{1/2}\cdot\lambda$$

όπου $C'_0 := \frac{C_0}{1-\delta-\theta} + \frac{\theta(1+\delta)}{(1-\delta-\theta)^2}$

Εφαρμόζοντας το δεύτερο κομμάτι του Λήμματος Α.1 και από την (Α.21) , συμπεραίνουμε ότι

$$\|h\|_{\ell_2} \le 2\|h_{01}\|_{\ell_2} + \frac{1+\delta}{1-\delta-\theta} S_0^{1/2} \cdot \lambda \le C_2 \cdot S_0^{1/2} \cdot \lambda$$

Και η ολοκλήρωση έρχεται από την σχέση (Α.15)

■

Πόρισμα Α.1 : Οριακοί Περιορισμοί (*Constrained thresholding*)

Έστω β να είναι S-αραιό έτσι ώστε

$$\|\beta\|_{\ell_2} \le \lambda \cdot S^{1/2}$$

για $\lambda > 0$. Τότε θα υπάρχει η ανάλυση $\beta = \beta' + \beta''$ έτσι ώστε

$$\|\beta'\|_{\ell_2} \le \frac{1+\delta}{1-\delta-\theta} \|\beta\|_{\ell_2}$$

$$\|\beta'\|_{\ell_1} \le \frac{1+\delta}{1-\delta-\theta} \frac{\|\beta\|_{\ell_2}^2}{\lambda}$$

και

$$\|X^* X\beta''\|_{\ell_\infty} < \frac{1-\delta^2}{1-\delta-\theta} \lambda$$

■

Απόδειξη Πορίσματος Α.1

Έστω $T := \{j : |\langle X\beta, X^j \rangle| \ge (1+\delta)\lambda\}$. Υποθέτουμε ότι $|T| \ge S$. Έπειτα μπορούμε να βρούμε το υποσύνολο T' του T με πληθικότητα $|T'| = S$. Τότε με χρήση της περιορισμένης ισομετρίας θα έχουμε

$$(1+\delta)^2 \lambda^2 S \le \sum_{j \in T'} |\langle X\beta, X^j \rangle|^2 \le (1+\delta)\|X\beta\|_{\ell_2}^2 \le (1+\delta)^2 \|\beta\|_{\ell_2}^2,$$

που έρχεται σε αντίθεση με την υπόθεση μας και άρα θα έχουμε $|T| < S$. Εφαρμόζοντας και πάλι την περιορισμένη ισομετρία καταλήγουμε στην ακόλουθη σχέση

$$(1+\delta)^2 \lambda^2 |T| \le \sum_{j \in T'} |\langle X\beta, X^j \rangle|^2 \le (1+\delta)^2 \|\beta\|_{\ell_2}^2$$

και επομένως

$$|T| \le S := \frac{\|\beta\|_{\ell_2}^2}{\lambda^2}.$$

Αφού ορίσουμε $c_j := \langle X\beta, X^j \rangle$, μπορούμε να βρούμε το β' έτσι ώστε

$$\langle X\beta', X^j \rangle = \langle X\beta, X^j \rangle \quad \text{για όλα τα } j \in T,$$

$$\|\beta'\|_{\ell_2} \le \frac{1+\delta}{1-\delta-\theta} \|\beta\|_{\ell_2},$$

$$\|\beta'\|_{\ell_1} \le \frac{(1+\delta)\sqrt{S}}{1-\delta-\theta} \|\beta\|_{\ell_2} = \frac{(1+\delta)}{1-\delta-\theta} \frac{\|\beta\|_{\ell_2}^2}{\lambda}$$

και

$$\left|\langle X\beta', X^j \rangle\right| \le \frac{\theta(1+\delta)}{(1-\delta-\theta)\sqrt{S}} \|\beta\|_{\ell_2} \text{ για όλα τα } \quad j \notin T$$

Από τον ορισμό του S, θα έχουμε

$$\left|\langle X\beta', X^j \rangle\right| \le \frac{\theta}{1-\delta-\theta}(1+\delta)\lambda \text{ για όλα τα } j \notin T$$

Επίσης, από τον ορισμό του T, θα έχουμε

$$\left|\langle X\beta, X^j \rangle\right| < (1+\delta)\lambda \text{ για όλα τα } j \notin T$$

Αφού ορίσουμε $\beta'' = \beta - \beta'$ καταλήγουμε στο αποτέλεσμα.

■

Α.2 Αποδείξεις Κεφαλαίου 6

Α.2.1 Απόδειξη Εξίσωσης (6.7)

$$\begin{aligned} \frac{1}{n}\left\|\hat{X}'\left(\hat{Y} - \hat{X}\hat{\beta}\right)\right\|_\infty &= \frac{1}{n}\left\|\hat{X}'\left[(I-K)(X\beta + g(T) + \epsilon) - \hat{X}\hat{\beta}\right]\right\|_\infty \\ &= \frac{1}{n}\left\|\hat{X}'\left[\hat{X}\beta + \hat{g} + (I-K)\epsilon - \hat{X}\hat{\beta}\right]\right\|_\infty \qquad (A.23) \\ &= \left\|\frac{1}{n}\hat{X}'\hat{X}\left(\beta - \hat{\beta}\right) + \frac{1}{n}\hat{X}'\hat{g} + \frac{1}{n}\hat{X}'(I-K)\epsilon\right\|_\infty \end{aligned}$$

Γνωρίζουμε ότι $\frac{1}{n}\hat{X}'\hat{X} \to \Sigma$ από τον Speckman. Έπειτα εξετάζουμε την i-οστή συνιστώσα των δύο τελευταίων όρων, $\frac{1}{n}\hat{X}'\hat{g} + \frac{1}{n}\hat{X}'(I-K)\epsilon$, δηλαδή την ποσότητα $\frac{1}{n}\left[\hat{X}^{i'}\hat{g} + \hat{X}^{i'}(I-K)\epsilon\right]$, όπου $\hat{X}^i = \left(\hat{X}_{1i}, \hat{X}_{2i} \dots, \hat{X}_{ni}\right)'$ είναι η i-οστή στήλη του πίνακα $\hat{X}$.

$$\frac{1}{n}\left[\hat{X}^{i'}\hat{g} + \hat{X}^{i'}(I-K)\epsilon\right] = \frac{1}{n}\left\{\left[\hat{q}^i + (I-K)\eta^i\right]'\hat{g} + \left[\hat{q}^i + (I-K)\eta^i\right]'(I-K)\epsilon\right\}$$

$$= \frac{1}{n}[\hat{q}^{i'}\hat{g} + \eta^{i'}\hat{g} - \left(K\eta^i\right)'\hat{g} + \hat{q}^{i'}\epsilon + \eta^{i'}\epsilon - \left(K\eta^i\right)'\epsilon - \hat{q}^{i'}K\epsilon - \eta^{i'} + \left(K\eta^i\right)'K\epsilon] \quad (A.24)$$

όπου $\hat{q}^i = \left(\hat{q}_i(T_1), \hat{q}_i(T_2), \dots, \hat{q}_i(T_n)\right)'$και $\eta^i = (\eta_{1i}, \eta_{2i}, \dots, \eta_{ni})'$ είναι η i-οστή στήλη του $\hat{q}$ και η αντίστοιχα. Κατόπιν θα δείξουμε ότι όλοι οι όροι είναι της τάξης $o_p(1)$ υπό της συνθήκες κανονικότητας (α)-(η). Από την συνθήκη $(\sigma\tau)$ συνεπάγεται ότι $\left\|\hat{q}^i\right\| = O_p\left(n^{1/10}\right)$, και άρα θα έχουμε $\frac{1}{n}\hat{q}^{i'}\hat{g} = O_p\left(n^{-4/5}\right) = o_p(1)$. Εξαιτίας του ότι ισχύουν οι ακόλουθες εξισώσεις, $E\left(\left\|K\eta^i\right\|^2\right) = E\left(E\left(\left\|K\eta^i\right\|^2|T\right)\right) = \sigma_{ii}E\left(trace(K'K)\right) = O\left(n^{1/5}\right)$,

$\left\|K\eta^i\right\| = O_p\left(n^{1/10}\right)$

$E(\|K\epsilon\|^2) = \sigma^2 E\left(trace(K'K)\right) = O\left(n^{1/5}\right)$, $\|K\epsilon\| = O_p\left(n^{1/10}\right)$

$E\left(\eta^{i'}\hat{g}\right) = 0$, $\mathrm{Var}\left(\eta^{i'}\hat{g}\right) = \sigma_{ii}E\|\hat{g}\|^2 = O\left(n^{1/5}\right)$, $\eta^{i'}\hat{g} = O_p\left(n^{1/10}\right)$

$E\left(q^{i'}\epsilon\right) = 0$, $\mathrm{Var}\left(q^{i'}\epsilon\right) = \sigma^2 E\left\|\hat{q}^i\right\|^2 = O\left(n^{1/5}\right)$, $\hat{q}^{i'}\epsilon = O_p\left(n^{1/10}\right)$

$E\left(\eta^{i'}\epsilon\right) = 0$, $\mathrm{Var}\left(\eta^{i'}\epsilon\right) = n\sigma^2\sigma_{ii}$, $\eta^{i'}\epsilon = O_p\left(n^{1/2}\right)$

το $\left(K\eta^i\right)'\hat{g}$ κυριαρχείται από $\left\|K\eta^i\right\| \cdot \|\hat{g}\| = O_p\left(n^{1/10}n^{1/10}\right) = O_p\left(n^{1/5}\right)$

το $\left(K\eta^i\right)'\epsilon$ κυριαρχείται από $\left\|K\eta^i\right\| \cdot \|\epsilon\| = O_p\left(n^{1/10}n^{1/2}\right) = O_p\left(n^{3/5}\right)$

το $\hat{q}^{i'}K\epsilon$ κυριαρχείται από $\left\|\hat{q}^i\right\| \cdot \|K\epsilon\| = O_p\left(n^{1/10}n^{1/10}\right) = O_p\left(n^{1/5}\right)\left(K\eta^i\right)'$

το $\left(K\eta^i\right)'\epsilon K\epsilon$ κυριαρχείται από $\left\|K\eta^i\right\| \cdot \|K\epsilon\| = O_p\left(n^{1/10}n^{1/10}\right) = O_p\left(n^{1/5}\right)$

, όλοι οι όροι της εξίσωσης (A.24) είναι της τάξεως $o_p(1)$, δηλαδή

$$\frac{1}{n}\hat{X}'\hat{g} + \frac{1}{n}\hat{X}'(I-K)\epsilon = o_p(1) \qquad (A.25)$$

Και άρα η εξίσωση (6.7) ισχύει.

■

Α.2.2 Απόδειξη Θεωρήματος 6.1

Για την περίπτωση $\alpha = \infty$ τα αποτελέσματα είναι προφανή από την (6.8). Υποθέτουμε ότι $\alpha < \infty$, και $Mat(p,R)$ να είναι η συλλογή όλων των $p \times p$ πινακών με πραγματικά στοιχεία και ορίζουμε $W = \{Q \in Mat(p,R)$, όπου Q να είναι θετικά ορισμένη και η Q δεν είναι παράλληλη με την ℓ^1-μπάλα$\}$.

Ορίζουμε τη συνάρτηση $F : Mat(p,R) \times R^p \times R \to R^p \cup \{\infty\}$ με

$$F(Q,\upsilon,\lambda) = \begin{cases} \tilde{\beta}(Q,\upsilon,\lambda), \varepsilon\acute{\alpha}\nu\ Q \in W \\ \\ \infty, \alpha\lambda\lambda\iota\acute{\omega}\varsigma \end{cases}$$

Όπου $Q \in W, \tilde{\beta}(Q,\upsilon,\lambda)$ είναι η μοναδική λύση στο πρόβλημα βελτιστοποίησης

$$\min_{\tilde{\beta}\in R^p}\left(\left\|\tilde{\beta}\right\|_1\right) \text{ έτσι ώστε} \left\|Q\left(\beta-\tilde{\beta}\right)+\upsilon\right\|_\infty \le \lambda \qquad (A.26)$$

Ακολούθως, θα δείξουμε ότι η συνάρτηση F είναι συνεχής στο $W \times R \times R$. Υποθέτουμε ότι $Q_n \in W,\ \upsilon_n \in R^p$ και $\lambda_n \in R, n = 1,2,\ldots$, είναι συγκλίνουσες ακολουθίες με $Q_n \to Q \in W$, $\upsilon_n \to \upsilon$ και $\lambda_n \to \lambda$. Έστω $\{\beta_n\}_n^\infty = 1$ είναι μια ακολουθία τέτοια ώστε $\beta_n \to \tilde{\beta} = \tilde{\beta}(Q,\upsilon,\lambda)$ και ικανοποιεί την ακόλουθη ανισότητα $\|Q_n(\beta-\beta_n)+\upsilon_n\|_\infty \le \lambda_n$. Για το πρόβλημα βελτιστοποίησης

$$\min_{\tilde{\beta}_n\in R^p}\left(\left\|\tilde{\beta}_n\right\|_1\right) \text{ τέτοια ώστε } \left\|Q_n\left(\beta-\tilde{\beta}_n\right)+\upsilon_n\right\|_\infty \le \lambda_n \qquad (A.27)$$

μπορούμε να πάρουμε ότι $\tilde{\beta}_n = \tilde{\beta}(Q_n,\upsilon_n,\lambda_n), n = 1,2,\ldots$ είναι λύσεις του (Α.27) και περιέχονται όλες σε ένα συμπαγές σύνολο από τον ορισμό της συνάρτησης F. Τώρα υποθέτουμε ότι $\tilde{\beta}_{n_k} \to \tilde{\beta}_0$ για ορισμένες ακολουθίες $\{n_k\}_{k=1}^\infty$ και $\tilde{\beta}_0 \in R^p$. Τότε $\tilde{\beta}_0$ είναι εφικτό για την (Α.26) και

$$\|\tilde{\beta}_0\|_1 = \lim_{k\to\infty}\|\tilde{\beta}_{n_k}\|_1 \le \lim_{k\to\infty}\|\beta_{n_k}\| = \|\tilde{\beta}\|_1$$

Εξαιτίας της μοναδικότητας και την βελτιστότητα της $\tilde{\beta}$, προκύπτει ότι $\tilde{\beta}_0 = \tilde{\beta}$. Αυτό συνεπάγεται ότι η F είναι συνεχής στο $W \times R^p \times R$. Τώρα έστω το $\hat{\beta}$ να αποτελεί λύση του Dantzig selector, τότε από τον ορισμό της συνάρτησης F και της εξίσωσης (Α.23) θα έχουμε

$$\hat{\beta} = F\left(\frac{1}{n}\hat{X}'\hat{X}, \frac{1}{n}\hat{X}'[\hat{g} + (I-K)\epsilon], \frac{1}{n}\lambda\sigma\right).$$

Λόγω του ότι ισχύουν $\Sigma \epsilon W, \frac{1}{n}\hat{X}'\hat{X} \to \Sigma, P\left(\frac{1}{n}\hat{X}'\hat{X} \epsilon W\right) \to 1, \frac{1}{n}\hat{X}'[\hat{g} + (I-K)\epsilon] = o_p(1)$ από την (Α.25) παίρνουμε $\hat{\beta} \xrightarrow{P} \beta_0$ και καταλήγουμε στο αποτέλεσμα.

■

Α.2.3 Απόδειξη Θεωρήματος 6.2

Το σημαντικό κομμάτι που μας ενδιαφέρει να αποδείξουμε για αυτό το θεώρημα είναι ότι $\sqrt{n}\left(\hat{\beta}_{C^*} - \beta_{C^*}\right) \xrightarrow{D} N\left(0, \sigma^2\Sigma_{C^*,C^*}\right)$. Υποθέτουμε ότι $\frac{\lambda\sigma}{\sqrt{n}}w_j \xrightarrow{P} \infty$ για $j \notin C^*$, και $\lambda\sigma w_j = o_p\left(\sqrt{n}\right)$ για $j \in C^*$, $C = \{j, \hat{\beta}_j \neq 0\}$. Τότε θα έχουμε ότι

$$\hat{\beta}_{C^*} = \left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{Y} - \lambda\sigma\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}W_{C^*,C^*r}$$

για $C = C^*$ και όπου $r \in R^{|C^*|}$, $\|r\|_\infty \le 1$. Τότε έπεται ότι:

$$\sqrt{n}\left(\hat{\beta}_{C^*} - \beta_{C^*}\right) = \sqrt{n}\left[\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{Y} - \lambda\sigma\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}W_{C^*,C^*r} - \beta_{C^*}\right]$$

$$= \sqrt{n}\left[\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{Y} - \beta_{C^*}\right] + \frac{1}{\sqrt{n}}\lambda\sigma\left(\Sigma^{-1}_{C^*,C^*} + o_p(1)\right)W_{C^*,C^*r}$$

$$= \sqrt{n}\left[\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\left(\hat{X}_{C^*}\beta_{C^*} + \hat{g} + (I-K)\epsilon\right) - \beta_{C^*}\right] + o_p(1)$$

$$= \sqrt{n}\left[\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{g} + \left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}(I-K)\epsilon\right] + o_p(1)$$

Γνωρίζουμε ότι $\hat{X}'_{C^*}\hat{g} = O_p\left(n^{1/5}\right)$ από την απόδειξη της εξίσωσης (6.7) υπό τις συνθήκες κανονικότητας (α)-(η), τότε θα έχουμε

$$\sqrt{n}\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{g} = \frac{1}{\sqrt{x}}\left(\frac{1}{n}\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}\hat{g} = O_p\left(n^{-3/10}\right) = o_p(1).$$

Έπεται ότι

$$\sqrt{n}\left(\hat{\beta}_{C^*} - \beta_{C^*}\right) = \sqrt{n}\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}(I-K)\epsilon + o_p(1)$$

Με βάση τον Speckman (1998) προκύπτει το ζητούμενο

$$\sqrt{n}\left(\hat{X}'_{C^*}\hat{X}_{C^*}\right)^{-1}\hat{X}'_{C^*}(I-K)\epsilon \xrightarrow{D} N\left(0, \sigma^2\Sigma_{C^*,C^*}\right)$$

■

ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] Akaike, H. (1974): "*A new look at the statistical model identification*," IEEE Transactions on Automatic Control.

[2] Anderson, T. W., and H. Rubin (1949): "*Estimation of the Parameters of Single Equation in a Complete System of Stochastic Equations*," Annals of Mathematical Statistics.

[3] Angrist, J., V. Chernozhukov, and I. Fernandez-Val (2006): "*Quantile Regression under Misspecification, with an Application to the U.S. Wage Structure*," Econometrica, 74(2).

[4] Belloni, A., and V. Chernozhukov (2011b): "*High Dimensional Sparse Econometric Models: An Introduction*," Inverse problems and high dimensional estimation - Stats in the Chateau summer school in econometrics and statistics, 2009, Springer Lecture Notes in Statistics – Proceedings.

[5] Belloni, A., and V. Chernozhukov (2011c): "*Least Squares After Model Selection in High-dimensional Sparse Models*," forthcoming Bernoulli.

[6] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "*Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain*," Preprint, ArXiv.

[7] Belloni, A., V. Chernozhukov, and C. Hansen (2010): "*LASSO Methods for Gaussian Instrumental Variables Models*," Preprint, ArXiv.

[8] Belloni, A., V. Chernozhukov, and L. Wang (2010): "*Square-Root-LASSO: Pivotal Recovery of Nonpara-metric Regression Functions via Conic Programming*," Preprint, ArXiv.

[9] Belloni,A. and Chernozhukov, and Hansen,C. (2011):"*Inference for High-dimensional sparse Econometric models*".

[10] Belloni,A. and Chernozhukov,V.: "*High Dimensional Sparse Econometric Models: An introduction*".

[11] Bernanke B, Boivin J, Eliasz PS (2005): " *Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach*," The Quarterly Journal of Economics. 120(1).

[12] Bickel, P. J., Y. Ritov, and A. B. Tsybakov (2009): "*Simultaneous analysis of Lasso and Dantzig selector*,"Annals of Statistics, 37(4).

[13] Boyd, S. and Vandenberghe L. (2004): "*Convex Optimization*". Cambridge University Press.

[14] Breiman,L. (1995): "*Better Subset Regression Using the Nonnegative Garrote*".

[15] Candès, E. J. and Tao, T. (2005): "*Decoding by linear programming*". IEEE Trans. Inform. Theory 51.

[16] Candès, E. J. and Tao, T. (2006): "*Near optimal signal recovery from random projections: universal encoding strategies?*" To appear in IEEE Trans. Inform. Theory 52.

[17] Candès, E. J., Romberg, J. and Tao, T. (2006): "*Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information*". IEEE Trans. Inform. Theory 52.

[18] Candès, E., and T. Tao (2007): "*The Dantzig selector: statistical estimation when p is much larger than n*,"Ann. Statist., 35(6).

[19] Chamberlain, G. (1987): "*Asymptotic Efficiency in Estimation with Conditional Moment Restrictions*," Jour-nal of Econometrics, 34.

[20] Chen, X. (2007): "*Large Sample Sieve Estimation of Semi-Nonparametric Models*," Handbook of Econometrics, 6.

[21] Chernozhukov, V. (2009): "*High-Dimensional Sparse Econometric Models*," (Lecture notes) Stats in the Chateau.

[22] Chernozhukov, V., and C. Hansen (2008a): "*Instrumental Variable Quantile Regression: A Robust Inference Approach*," Journal of Econometrics, 142.

[23] Craven P., Wahba G. (1979): "*Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation*". Numer Math 31.

[24] Dicker L, Lin X (2009): "A large sample analysis of the Dantzig selector and extensions".

[25] Donoho, D. L. and Johnstone, I. M. (1994): "*Ideal denoising in an orthonormal basis chosen from a library of bases*". C. R. Acad. Sci. Paris S´er. I Math. 319.

[26] Donoho, D. L. and Johnstone, I. M. (1994): "*Ideal Spatial Adaptation by Wavelet Shrinkage*". Biometrika 81 .

[27] Fan J, Li R (2004): "*New Estimation and Model Selection Procedures for Semiparametric Modeling in Longitudinal Data Analysis*". J Am Stat Assoc 99.

[28] Fan,Y., and Lv,J. (2009): "*A unified approach to model selection and sparse recovery using regularized least squares*".

[29] Foster, D. P. and George, E. I. (1994): "*The Risk Inflation Criterion for Multiple Regression*". Ann. Statist. 22.

[30] Frank,I. and Friedman,J. (1993): "*A Statistical View of Some Chemometrics Regression Tools*".

[31] Gautier, E., and A. Tsybakov (2011): "*High-dimensional Instrumental Variables Rergession and Confidence Sets*," Preprint, ArXiv.

[32] James GM, Radchenkob P, Lv JC (2009): "*DASSO: connections between the Dantzig selector and LASSO*". J R Stat Soc Ser B 71.

[33] Koenker, R., and G. Bassett (1978): "*Regression Quantiles*," Econometrica, 46.

[34] Li, R., and Fan, J. (2009): "*Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties* ".

[35] Li,F. and Lin,L. and Su,Y. (2012): "*Variable selection and parameter estimation for partially linear models via Dantzig Selector*".

[36] Meinshausen, N., and B. Yu (2009): "*Lasso-type recovery of sparse representations for high-dimensional data*," Annals of Statistics, 37(1).

[37] Natarajan, B. K. (1995): "*Sparse approximate solutions to linear systems*". SIAM J. Comput. 24.

[38] Newey, W. K. (1997): "*Convergence Rates and Asymptotic Normality for Series Estimators*," Journal of Econometrics, 79 .

[39] Ruppert D, Sheather S.J., Wand M.P. (1995): "*An Effective Bandwidth Belector for Local Least Squares Regression*". J Am Stat Assoc 90 .

[40] Schwarz, G. (1978): "*Estimating the Dimension of a Model*," Annals of Statistics, 6(2).

[41] Sims C.A, (1980): "*Macroeconomics and Reality. Econometrica*". 48(1).

[42] Staiger, D., and J. H. Stock (1997): "*Instrumental Variables Regression with Weak Instruments*," Econometrica, 65.

[43] Stock J.H., Watson M.W. (2001) : "*Vector autoregressions: The Journal of Economic Perspectives*".15(4).

[44] Tibshirani, R. (1996): "*Regression shrinkage and selection via the Lasso*," J. oy. Statist. Soc. Ser. B, 58.

[45] van de Geer, S. A. (2008): "*High-dimensional generalized linear models and the lasso*," Annals of Statistics, 36(2).

[46] Wang H, Li R, Tsai C (2007): "*Tuning parameter selectors for the smoothly clipped absolute deviation method*". Biometrika 94.

[47] Wang Lie, Brown LD, Cai TT (2011): "*A difference based approach to the semiparametric partially linear model*" . Electron J Stat 5.

[48] Yatchew A (1997) An elementary estimator for the partially linear model. Econ Lett 57.

[49] Zhang, C.-H. (2010): "*Nearly unbiased variable selection under minimax concave penalty*," Ann. Statist., 38(2).

[50] Zou,H. and Hastie,T. (2005) : "*Regularization and variable selection via the elastic net*".